MAKSWELL

MAKing Sustainable development
and WELL-being frameworks work for policy analysis

www.makswell.eu

Horizon 2020 - Research and Innovation Framework Programme
Call: H2020-SC6-CO-CREATION-2017
Coordination and support actions (Coordinating actions)

Grant Agreement Number 770643

Work Package 3

Regional poverty measurement as a prototype for modern indicator methodology

Deliverable 3.2

Guidelines for best practices implementation for transferring methodology

September 2020

Destatis, Istat, Statistics Netherlands, Pisa University,
Southampton University, Trier University

This project has received funding from the European Union's Horizon 2020
research and innovation programme.

Deliverable 3.2

Guidelines for best practices implementation for transferring methodology

Authors

Destatis:
Thomas Zimmermann

Istat:
Federico Polidoro, Federico Di Leo, Massimo Fedeli

Statistics Netherland - CBS:
Joep Burger, Jan van den Brakel

University of Pisa - Dagum Center ASESD:
Monica Pratesi, Caterina Giusti, Stefano Marchetti, Luigi Biggeri, Gaia Bertarelli,
Francesco Schirripa Spagnolo, Tiziana Laureti, Ilaria Benedetti

University of Southampton:
Paul A. Smith, James Dawber, Nikos Tzavidis, Angela Luna

Office for National Statistics:
Jim O'Donoghue, Tanya Flower, Heledd Thomas

Freie Universität Berlin:
Nora Würz, Timo Schmid

Universität Trier:
Charlotte Articus, Jan Pablo Burgard, Christopher Caratiola, Hanna Dieckmann,
Florian Ertz, Joscha Krause, Ralf Münnich, Anna-Lena Wölwer

Summary

Today, *big data* is a buzz word. Although there have been attempts to properly define the term, a really universally accepted definition has not yet been given. Accordingly, many different types of data may be classified as *big data* or *new data*. These range from scanner data collected at retail outlets, through remote sensing data to mobile phone data. As the availability of such data increases, researchers try to make use of them by incorporating them into existing methods and developing new methods. These developments are also highly relevant for the estimation of well-being indicators, a core focus of the MAKSWELL project. The combination of new data sources and new or modified methods are promising especially where the estimation of well-being at a fine spatial resolution is concerned. While a comprehensive survey of the related literature and available data sets is out of the scope of this project, this deliverable collects a few (experimental) applications that shed a light on the potential benefits of these new approaches. Some drawbacks and practical implementation problems are addressed as well. Taken as a whole, the presented set of applications points to future research needs in the area and allows the derivation of some general best practice guidelines that can also inform other subject matter areas beside the measurement of poverty and well-being.

1

Introduction

The reduction of poverty is one of the EU priorities. To measure poverty and well-being adequately, the Statistics on Income and Living Conditions (SILC) was developed including a variety of indicators such as AROPE. European policies however do not only focus on countries but prominently target European regions; more than one third of the European Union's budget is devoted to its cohesion policy. Hence, an adequate statistical methodology has to be developed to enable an accurate regional measurement by indicators.

The aim of this deliverable is to provide an overview of big and new data use to improve the quality of regional indicators in the wide area of poverty and well-being measurement. The data sources discussed in this deliverable cover scanner and satellite data as two examples of big data. Mobile phone data still suffer from monopolies of the mobile phone provider such that they are either extremely expensive or not available in the necessary details, though first attempts especially for estimating and analyzing commuter behaviour are promising.

This deliverable is organised as follows. Section 2 covers some recent (experimental) applications that use new (i.e. non-traditional) data sources to estimate measures related to poverty at a fine spatial resolution. Subsection 2.a first introduces the concept of regional price indices and details their construction. It also discusses some of the issues involved in their estimation and touches on new data sources that could be used for such indices. After a discussion of variance and quality estimation for regional price indices, the former is demonstrated for the UK. Subsection 2.b gives an insight into recent work on the estimation of sub-national spatial consumer price indices using scanner data in Italy and sketches the impact of local cost-of-living differences on poverty incidence. After a short recapitulation of area-level models for small area estimation (SAE) , Subsection 2.c details work on the estimation of wealth at the level of Upazilas in Bangladesh using SAE and remote sensing covariates. Subsection 2.d describes an exploratory case study on the potential of employing remote sensing data for the disaggregation or downscaling of official statistics on poverty and income using remote sensing data. Data for three large cities in the Netherlands has been used. Subsection 2.e concludes Section 2 with an experimental application of SAE to produce regionally disaggregated estimates of the number of people earning the general statutory minimum wage in Germany, where covariates have been measured with sampling error. Section 3 briefly outlines future research needs related to the estimation of regional measures of well-being. Additionally, some best practice guidelines for future applications are given. Section 4 concludes the deliverable.

2

Applications: Measuring poverty and wealth using new data sources

2.a. Regional price indices

2.a.1. Introduction

Local indicators of wellbeing cover a wide range of types of indicators, including a range of economic indicators which are measured in monetary terms, or compiled from sources which are collected in this way and then processed further. This typically requires some adjustment to deal with changes in the value of money, a process of deflation which is well established, but which developed over a long period (O'Neill et al., 2017). This is accomplished nationally through the calculation of price indices which measure the change in the value of money, with a range of different price indicators typically being available for different components of change.

National Statistical Institutes (NSIs) have well-developed processes to collect the data on prices and quantities needed to calculate these price indices at a national level, and the size of the data collection operation is tailored to the production of a reasonable quality at this level. However, there has been a long interest in the way that prices vary across regions within countries, and this can be measured in two ways. The most straightforward one is to measure the level of prices in different regions at the same time, and this is the basis of Purchasing Power Parity (PPP) measurements, which are an established component of price measurement in the European Union (EU) (Eurostat and OECD, 2012). PPPs give rise to spatial indices which compare a regional price level to a comparison region (or some function of them such as the mean). They measure the price levels, and are therefore not usable for deflation to account for changes in prices regionally, though they can be used to adjust for the differences in prices (or the purchasing power of money) in different regions at a specific time, as in the examples in 2.b .

The regular PPP price collections may take place in a country's capital city, since the items in the basket, which include items characteristic of expenditure in other countries, are not always widely available elsewhere. Periodically (at least every six years under EU regulation (EC) No 1445/2007) a regional data collection is used to calculate spatial adjustment factors, which adjust the capital city collection to be representative nationally. In the UK, these data are used to produce regional price levels as a by-product (Baran and O'Donoghue, 2002, Wingfield et al., 2005, ONS, 2011). However, the changes in the weights between consecutive periods are too great to enable even an approximation of regional *temporal* inflation.

However, there is a continuing interest in regional temporal price indices which would provide the basic materials to enable regional price deflation. In some countries, there is a regional component to the way price indices, particularly consumer prices indices, are constructed, for example in the USA, or in Japan. In this case regional prices are formed naturally as part of the aggregation process. In other countries such as the UK, there is no regional stage in the aggregation, and many people have therefore made attempts to construct approximate regional prices by using the data which are available (see chapter 2.a.4). The construction of official regional temporal consumer price indices has been considered in the UK since the 1970s, but the data collection requirement has always been beyond the resources available, and the use of available data has been rejected on the grounds that it would produce regional indices of insufficient quality for use.

Here, we consider several stages in the development of regional CPIs, focusing on the UK situation as a case study. Chapter 2.a.2 sets out the conceptual framework for regional temporal consumer price indices, so that it is clear what target we are attempting to measure, even in situations where we have to fall back on approximations of the ideal measures. Chapter 2.a.4 reports attempts to construct regional CPIs for the UK, using existing price and household expenditure collections. This includes the investigation of small area estimators to assist in the construction of regional baskets and regional expenditure weights from the existing household survey data. It is important that any regional CPI which is produced, can be accompanied by an assessment of its quality, so that users can know how much confidence to place in measures derived with this information. In this chapter, we use an approximate measure based on the standard deviation of the first differences in the series. Unfortunately, the development of error measures for CPIs has proven to be very challenging, and while there is research in this area, it is scattered. Chapter 2.a.5 therefore presents a detailed review of the approaches to quality measurement in consumer price indices. Only in the USA, this kind of calculation is undertaken regularly; the use of regional components in the aggregation in the US CPI means that variances for regional CPIs are provided as a by-product of this processing. However, most work has considered only the quality of the national CPI. Chapter 2.a.6 presents recent work to estimate sampling variances for the UK CPI at national level which will provide a comparator for regional CPIs, although the work has not yet been extended to calculations at this level. Further conclusions, which draw together the links between these strands of work, generalises the lessons from the UK to other countries and sets out some directions for future research is presented in chapter 3.a.

2.a.2. Conceptual framework for regional price indices

2.a.2.1. Introduction

In order to have a sound basis for the methodology of regional price indices, it is important to set out the differences in the conceptual framework from the calculation of a national index. This chapter starts this process for consumer prices by addressing the elements that are related to regions, and highlighting issues with regional boundaries (which may be considered by analogy with national boundaries, but are typically much more porous to trade). It sets out target concepts which we ideally would like to cover, though in most cases, there will not be a data source which follows this definition sufficiently precise. So, we will consider the options for data sources, how close they come to the concept required, and therefore, what the best approximation to the ideal definition of a regional price index is.

Because there will in general be no ideal data source, it will be very difficult to quantify the approximation error in the calculation of regional price indices, but we will attempt a broad brush description of these errors, and make an assessment of their effects on the interpretation of the experimental indices.

2.a.2.1.1. Basic structure of consumer price indices

The UK CPI is a fixed basket index, where a range of goods and services (the ‘basket’) is priced each month, and the expenditure shares on items in the basket are used to weight the price information together. The expenditure shares derive primarily from the Living Costs and Food Survey (LCF), with adjustments for coverage and for other data sources related to specific items, and balanced through the National Accounts. The balanced expenditure shares are price updated, so that they are as close as possible to the required concept at the national level (see ONS (2019) for full details). The basket is updated each year, and the 13 month-long segments of monthly prices are joined together by chainlinking. The current implementation of the UK CPI involves a double chain link in December and January, which (since March 2017) is price updated in a manner which makes it consistent with a single chain link (from February).

2.a.2.1.2. Price levels or price changes

Regional price levels have been produced a few times in the UK (Baran and O’Donoghue, 2002, ONS, 2011, 2018b), and involve comparing price levels across regions – a spatial index. This was also the focus of other reviews of regional prices (Department of Employment, 1971, Fenwick and O’Donoghue, 2003), and shows the relative costs of a fixed basket of goods and services in different regions. Many prices are expected to vary rather little by region, and some prices which are set more or less nationally may not vary at all (eg mail order prices). A few commodities may behave very differently in different regions, and chief amongst these is housing. This would not be needed for a CPI-style index (which does not include housing costs), but would be very important (and a major component of regional differences) in a CPIH style index (which does).

The existing publications on Relative Regional Consumer Price Levels (RRCPLs) are derived from the information used in the 6-yearly benchmarking of the Purchasing Power Parity (PPP) series. PPP price quotes are collected only from the capital city (mainly on cost grounds, but also because they include items for which quotes are difficult to source in every region), and every six years a separate exercise is undertaken to calculate spatial adjustment factors, used to rescale the capital city PPP to be representative of the prices across the country. The information used in this rescaling also gives estimates of regional price levels for a fixed, national basket of goods and services, excluding housing costs (neither owner occupiers costs nor property rents are included). They use regional weights for the aggregation of the 111 basic headings, derived from the Living Costs and Food Survey, reflecting at least part of the differences in expenditure patterns by region.

The regional price levels are produced by imposing transitivity on the calculated indices, to produce a set of regional indices which can be satisfactorily compared (for an overview see (ONS, 2011, Annex 2)).

An alternative approach, also considered by the RPI Advisory Committee (Department of Employment, 1971) is to produce a temporal price index in each region. In effect, this means replicating the production of the CPIH in each region, and the conceptual challenges below apply more to this type of index. The RPI Advisory Committee thought that such indices could be produced annually (for London, Wales, Scotland and N Ireland) with some additional price collection.

In chapter 2.a.4 we take the challenge to be one of constructing experimental regional temporal price indices, that is measuring the difference in the rate of inflation in each region. It should be made clear that such indices will not produce information that is suitable for comparing price levels between regions. In order to have both regional inflation and regional price levels, it will generally be necessary to produce two types of indices. It is theoretically possible to produce a spatiotemporal index to allow simultaneous assessment of changes by region and across time, but the requirement for transitivity would mean that it would need to be revised each period (see also section 3.a.2.1).

2.a.2.2. Regional basket

The starting point for a regional price index should be the regional basket of goods and services. The national basket in the UK is derived from the Living Costs and Food Survey (LCF), supplemented by other sources for specific products. The products (COICOP4 up to Jan 2017 and COICOP5 from the Feb 2017 index) with the largest proportions of expenditure form the basket, though there are adjustments at the margins based on whether product sales are growing or shrinking, and for other relevant factors (Gooding, 2016). The maximum threshold is set at 1 part per thousand (ppt) according to the EU regulation 1687/98 on HICP. The UK implementation however uses a lower threshold such that categories with a minimum national expenditure of £400m pa are always included unless they are satisfactorily represented by other items, and items with expenditure less than £100m are normally not included (Gooding 2016). Commodities with expenditures between these values are reviewed to consider whether their pattern of sales suggests that they are emerging and should be added to the basket or waning and should be removed from it.

Based on total household expenditure estimates from the LCF (2014-16), £100m amounts to approximately 0.14 ppt and £400m to 0.56 ppt. Although it is possible to produce regional thresholds by applying the proportional equivalents of the £100m and £400m values to the regional expenditure totals, this seems unnecessarily complicated. For the exploratory calculations in chapter 2.a.4 we propose therefore to take a strict application of a ppt rule, and suggest taking 0.5 ppt initially. A sensitivity analysis could be used to consider the effect of this value on the size of the basket and the accuracy of the quantities within it, including the sampling variation in quantities near the threshold. A case could be made for lowering the threshold to something more intermediate between the £100m and £400m equivalents, or indeed for raising it to the 1ppt of the EU regulation in order to simplify calculations and reduce the impact of large sampling variances.

The regional procedure in chapter 2.a.4 emulates the national one by using an expenditure threshold. A process of deciding on inclusion/exclusion for borderline cases could be undertaken, but here we use the estimated proportions without manual intervention to define the baskets.

The LCF remains more or less the only source of information on consumers' purchases with sufficient detail to produce this information. The sample size is naturally much smaller for regions ($c.6000/14 \approx 400$), so the accuracy of the information on the basket will be reduced.

It is still likely that sample sizes will be sufficient for direct estimation of expenditure patterns. Some care may be needed with Northern Ireland, to check whether the variance is unusually large. From survey year 2016/17 NISRA have boosted the LCF in Northern Ireland to around 500 households

Table 2.a.1.: Regional sample sizes of households in the 2013 LCF

Government Office Region modified	No. of sample households
North East	251
North West and Merseyside	585
Yorkshire and the Humber	462
East Midlands	424
West Midlands	526
Eastern	497
London	480
South East	681
South West	429
Wales	246
Scotland	412
Northern Ireland	151
Total	5,144

(from 150 households). In the past, three years of LCF data have been grouped when producing outputs. Continuing this idea of merging years (even with 500 households) would seem to bring the most stability to estimates, although specific small area methods might be more targeted. If there were to be consideration of still smaller areas, then there would need to be an assessment of whether small area methods might be needed. These are discussed further below.

It is quite likely that some products included in the regional baskets will be insufficiently important nationally to be included in CPI price collection. For this exploratory calculation we will not attempt to provide price quotes for such products, and will treat them as adding their weight to the nearest similar product group.

There is a need for longitudinal consistency for baskets. The basket quantities should not fluctuate from year to year largely through sampling variation. One approach to this is to use information from several years of the LCF in calculating the baskets. More explicit smoothing is also possible. Any of these approaches risk introducing a small bias through lack of response to changes (since smoothing reduces responsiveness to change).

2.a.2.3. Regional weights

Section weights for the CPI are derived from household final consumption expenditure (HFCE), with few exceptions. ONS (2016b) discussed the feasibility of calculating regional HFCE, and published experimental HFCE at the regional level in ONS (2018a).

Conceptually, we should use these regional weights, to be consistent with CPIH construction, but we can approximate them by using the LCF regional expenditure patterns to break the weights into regional pieces, and then rescaling the sum of weights in each region to 1000, and we have used this approach in chapter 2.a.4. So

$$\hat{w}_{rk} = w_{nk} \frac{w_{rc}}{w_{nc}} \tag{2.a.1}$$

where w_{nk} and $\hat{w}_{rk}$ are respectively the national and estimated regional weight for item k, and w_{nc} and w_{rc} are respectively the national and regional estimates of household spending on class c derived from the LCF (and related sources in a few cases). The $\hat{w}_{rk}$ are then rescaled so that they sum up to 1000.

CPIH is not weighted only at the section level, however. More detail is obtained for commodity groups based on LCF and related sources below the level of HFCE. These too can be calculated by a further application of equations 2.a.1 which will subdivide each of the regional weights based on LCF expenditure patterns.

2.a.2.4. Business-based estimation of regional consumer spending

The regional baskets derived from consumer spending information from the LCF and associated sources have several advantages in the detail of the commodity breakdown which is available. However, they also have challenges because of cross-boundary expenditure, non-geographic expenditure and all the issues that arise from collecting data with surveys and diaries (non-response, measurement error, recall bias, etc.).

An alternative is to gather the information from the businesses. This overcomes some of the disadvantages of the consumer-based approach, but has compensating disadvantages. It avoids the boundary issues where consumers spend in regions other than where they live, because the businesses know the location of the spending. We still have some challenges around how to cope with non-geographic spending, such as mail order (but see below). Responding to business surveys is compulsory, which means that there is less of an issue with response (and none at all with diaries, though for some of the very smallest businesses, there may be no computerised records, giving a parallel data retrieval issue).

This generates a statistical challenge which would benefit from further research – is imputation for missing values of commodity sales in business surveys better than the equivalent imputation in social surveys? Business surveys have fewer predictor variables, and we might expect that there is less correlation between them and commodity sales. But the predictor variables are substantially complete, because they come from administrative data systems, in contrast to social survey predictors.

The largest businesses are crucial to making good quality estimates. Mostly, surveys do not collect a cross-classification of sales by region and product, so there is a choice between a very detailed data collection and an estimation challenge to produce satisfactory basket and weight type data from currently collected business survey data.

ONS is investigating the production of regional household expenditure data from a balancing process similar to that used in the national element of the national accounts. Such a process, producing balanced estimates from all the available inputs, may well be the best long-term strategy, since it uses all the sources of data and accounts for their relative accuracy in the balancing process.

2.a.2.4.1. Alternative data sources

The idea of collecting data from businesses may be sound in the case that alternative ("big") data sources become available. Scanner data, or suitably calculated summaries, could identify detailed commodity by region breakdowns for geographical sales. The delivery address for non-geographic sales could be used as the basis of a similar breakdown, and although this may have some minor inaccuracies with respect to gifts etc., it would provide a sounder basis for detailed estimation. Some investigation of the selectivity of the data may be needed – it is anticipated that such data would be acquired from the largest businesses, and so might not represent the smallest ones. But a single-region assumption for the smallest businesses might well be satisfactory. Perhaps the biggest challenges are gaining access to the data, and processing it to produce the required summaries. (Further exploration of the use of alternative data sources in regional estimation follows in chapter 2.a.3.)

2.a.2.5. Boundary issues: What is the definition of 'region' that a regional CPI covers?

There are two competing ways to define a region for CPI purposes. The first is to use the region in which expenditure takes place, and the second is to use the 'usual residence' of the person making the spend. The CPI manual uses the former definition in the calculation of the national index, which should include the spend of foreign nationals visiting the UK. For some products, we would not expect there to be large differences, whichever definition is used. But for other products (particularly perhaps for larger or more expensive purchases), there may be a noticeable difference with consumers making trips for particular shopping purposes.

By analogy with the national index, we would want to use the region of expenditure in a regional CPI, so that it would reflect the prices and weights of spend in a region. However, we do not have consumers' expenditure broken down by region and by product (though it may be possible to construct a version just by region-based on regional accounts, see section 2.a.2.3. This could perhaps be used for benchmarking to make an adjustment).

There are some boundary issues for Northern Ireland with the Irish Republic. This is primarily for items like petrol, where lower Irish Republic taxation levels provide advantages for Northern Ireland residents buying there. This does not necessarily invalidate any analysis, but needs to be noted and factored in to procedures. It may require some estimate of cross-border trade.

Additionally, we need to define a suitable procedure for dealing with non-locational purchases, such as mail-order goods, or services provided over the internet. One simple approach is to take the national price index for mail order etc goods in this case. But if there are considerable differences in patterns of purchases by region, it may be better to construct bespoke indices of mail order etc goods for different regions, using price quotes only for the appropriate products.

This can be couched in terms of a data challenge: would it be reasonable to collect information on region of expenditure (and non-regional expenditure) from a diary-based household budget survey such as the LCF? In specific cases, it would be necessary to assess, whether this information was already potentially available, or whether it could be coded from receipts provided when the LCF diary is completed. Once this data is available, it would be possible to assess the difference in the basis of the regional definition, although there is a risk that the information will be insufficiently detailed.

Ultimately, a wide provision of scanner data would enable a much better assessment of differences caused by the way region is defined.

There is a benefit too in public understanding, since a ‘regional CPI’ intuitively feels like the inflation rate for prices in a region, not the rate experienced by people living in a region who may shop regularly outside it. Though it may be that the difference is small and too subtle for non-expert users.

The alternative option, of regional prices weighted according to the region of usual residence of the spender, is less like the concept of the CPI, but more practical in its application. It can be estimated directly from the LCF data, and does not have any definitional challenges for non-locational purchases. There is a small issue over whether overseas spending can be identified and excluded.

So, accepting that the concept should match the CPI, we nevertheless recommend using the region of usual residence of the spender for the exploration in chapter 2.a.4. This should drive the derivation of the regional weights. This does leave outstanding the question of what to do with expenditure of foreign visitors. These are covered in CPI, and an adjustment to the LCF data inputs is made for them in balancing the national accounts. The same adjustment process can be applied for a rescaling of LCF inputs, at least implicitly in the rescaling of equation 2.a.1. More sensitive approaches based on numbers (and possibly spend) of foreign visitors may be worth investigation.

2.a.2.6. Regional price quotes

Clearly, the price information is needed from the region to which the spending relates. Price collectors already geolocate their physical collections, and central shops also have prices defined by a particular shop within a region, so regional versions of all prices are available. Care needs to be taken with central pricing for goods which are not stocked locally – ideally these should be excluded from the prices available for an index (if they can be identified).

This presents a further data challenge: in central price collection, is it clear which products are available by region? For example there are “national” beers which are available in all main Tesco stores (e.g. Fursty Ferret), but there are also local beers which are only stocked in the region in which they are produced. Is it possible to distinguish that Cwrw Haf (a Welsh beer) is only for sale in Wales, based on central price data? And similarly for other regional products.

2.a.2.7. Consistency in aggregation for regional price indices

The approaches discussed and recommended here are targeted at getting the best estimate of regional price inflation. This means that the measures are not using average baskets (and therefore are not consistently measured) across regions. Even if they were averaged in this way, it would be a challenge to develop a method whereby the indices would sum up to the national index and still provide appropriate regional price inflation measures. The outlined approaches therefore do not offer any aggregation above regional level. While it is in principle possible to aggregate the indices, the result will not measure any sensible concept, and should not be interpreted.

2.a.3. Alternative data sources to support calculation of regional indices

The traditional approach to the calculation of price indices involves the collection of a detailed dataset of price quotes along with detailed information about the product (or service) that the quote belongs to. This detailed specification (Morgenstern, 1963, p. 185) is needed to allow the quality of the priced item to be monitored, and a suitable adjustment made if the quality changes in an important way . This means that price collection is a detailed operation, typically involving a large field force to collect at least some of the prices directly from outlets in sampled areas which represent the transactions occurring throughout the territory for which the price index is to be calculated.

At the lowest level, there is usually no weighting information, and the prices are combined using an appropriate *elementary aggregation* formula to produce a bottom-level index which can then be aggregated using weighting information in a Laspeyres-type index. The weights are traditionally derived principally from household budget (household expenditure) surveys, and may be combined with other sources within a national accounts framework to correct for misreporting of particular items (such as alcohol, where there is an alternative source from tax data) and discrepancies between multiple sources covering the same items.

Dividing the prices from traditional collection operations into regions naturally produces fewer price quotes in each region than nationally. Therefore, if the calculation of regional indices is based on these data, their accuracy will be poorer than the national level indices. This kind of approach is explored in chapter 2.a.4 which also considers model-based (small area) procedures to improve the estimates of expenditure weights so that regional aggregate indices are improved. However, price indices have also been an area with considerable development of alternative data sources, and in this chapter, we consider from a conceptual view the possibilities to use this information either as the basis for, or to improve the estimation of regional price indices.

2.a.3.1. Types of regional price indices

Section 2.a.2.1.2 explains the difference between temporal and spatial regional price indices. Both will be considered here. The work by UNIPI in chapter 2.b uses two versions of a spatial index which differ mainly in the data sources — purchasing power parity (PPP) spatial indices are derived from common products priced in each region, whereas spatial price indices (SPIs) are derived from representative, but not necessarily common, products. The procedures used in chapter 2.a.4 correspond with the second of these, with specific products not necessarily common between regions. Most examples of regional price indices are of the spatial type, because these require only periodic data collection of sufficient size at the regional level.

Temporal regional price indices are rarer, though some countries (eg the USA) do produce them. In general, these are only practical where the aggregation structure for the price index includes a regional component, such that the regional indices are produced during the regular processing of the index. One strategy for producing regular regional indices would therefore be to reorganise the aggregation structure to have this form. Most consumer price indices however have so much history and value to users in their existing forms that making such changes is not straightforward. However, there is an additional advantage that the temporal regional price indices produced so are consistent with the national index, because they are a component of its calculation. This is not so for temporal regional indices calculated by a separate aggregation (see section 2.a.2.7).

In theory it is possible to produce a set of spatiotemporal indices designed to meet both objectives simultaneously, but this leads to a different set of challenges, discussed in more detail in section 3.a.2.1.

2.a.3.2. Approaches to use of alternative data sources for prices

There are two principal sources of alternative data for prices - from scanner data provided specifically by businesses and by scraping from the websites of businesses selling to the public. Both sources require preprocessing before they can be used. Scanner data are typically easier to handle as they have a reference which can be used to identify a product, but this reference may persist even when the quality of the item changes. Detecting such a quality change may therefore not be straightforward, and some work is needed to understand how such changes can be identified and adjusted for in scanner data. An identified product can be classified, and this classification can persist as long as the product identifier is unchanged. It is however still expected that there will be a steady flow of products needing classification decisions each time new data are received.

Classification is a bigger challenge from information scraped from the web, which may not include all the detail that would ideally be required to make a classification decision, or to detect quality changes. In both scanner data and scraped data, there are very large numbers of products (compared with a traditional price collection) which need to be monitored and coded, and this has led to investigations involving machine learning to automatically classify all or a large proportion of prices, for example see Harms and Spinder (2019), Myklatun (2019), ONS (2020a).

Assuming that we have information from alternative data sources suitably cleaned and coded, there are still options for how the data can be used, as described below.

2.a.3.2.1. Directly as a source of prices

One option is to use the collected prices directly as the basis for constructing an index. Several countries have already introduced alternative data sources into their consumer price indices (eg Belgium (van Loon and Roels, 2018), Slovenia (Noč Razinger, 2018)), with some element of replacement of direct collections from stores, so that the price quotes from the alternative data sources are integrated into the standard processing of the index. This is the first stage in the development of price indices towards alternative data sources as described by Zhang (2018).

A second stage involves the use of different index formulae to deal with the large numbers of quotes, more or less rapidly changing products, and consequent lack of identifiable changes in prices. Payne (2017) shows the extent of churn in a set of price quote data for clothing. In these cases, a different approach to index aggregation is needed, and a range of methods has been considered, although it is not clear that any has currently solved the downward drift in prices where there is large churn in the products available. These indices seem to have better properties in dealing with more stable products, such as staple foods, and these are naturally the components of a price index where these data sources have been most widely used to date.

Clearly, once the methodology for using scanner data or web scraped data for the calculation of price indices at a national level has been worked out, it will not be particularly challenging to replicate this process at a regional level, and the calculation of regional temporal indices may become relatively straightforward. Despite the inclusions of these sources, no country is yet able to produce a CPI purely from such data sources, although the billion prices project (Cavallo and Rigobon, 2016) does manage a reasonable approximation to CPIs using only this type of data. Unfortunately, the methodological challenges of dealing with this data are still being tackled, so this situation is not yet in view.

The availability of scanner and web scraped data should also allow reasonably straightforward extensions to spatial price indices as used by UNIPI in chapter 2.b, which use whatever prices are available. But the stricter control of the products in a PPP index are likely to substantially reduce the numbers of quotes which are available from these alternative sources. Further experience is needed to see, whether it is possible to construct a reasonable PPP with such data.

2.a.3.2.2. To supplement direct price collection

Scanner data and, particularly, web-scraped data are both susceptible to lower quality than directly collected data, because it is harder to monitor quality changes and unusual price movements from only the available data. Mayhew and Clews (2016) show, how machine learning can be used to identify unusual movements, and these methods may improve the quality of alternative data sources as inputs for scanner data. While these methods mature, it would be possible to combine index estimates from the prices collected directly, with high quality, with index estimates from alternative data sources, with lower quality. This suggests some form of composite estimation which normally accounts for differences in variability in two sources estimating the same quantity. In this case, the evaluation of the quality is complicated by the difficulty of defining the sampling variance of the standard CPI (see chapters 2.a.5 and 2.a.6), and also because there is a potential for bias in the measurement, but which cannot be evaluated as there is no target statistic against which to compare. Further research would be needed to define satisfactory measures, probably a version of mean squared error, which could be used to combine statistics produced from the two different data collection sources.

Most combinations in use in national CPIs replace a portion of the collected data by alternative sources instead of combining the two sources.

If a national approach to composite estimation approach could be devised, it should again be relatively easy to extend it to regional indices; calculation of the required quality measures would allow a combination at this lower level of geography, where the sampling error in the controlled data collection, with a substantially reduced number of prices, would be offset by the larger number of prices from the alternative data source.

2.a.3.2.3. In models of price change

In chapter 2.a.4 below, we consider the use of model-based estimation (small area estimation) to provide improved estimates of expenditure weights for price indices. In principle, it should also be possible to build models which describe prices or price changes, and to use these in model-based approaches to improve estimates of prices or price changes. This is related to the idea of composite estimation using collected and scraped/scanner prices (See section 2.a.3.2.2). There are several conceptual challenges, however.

The first challenge is to gather suitable predictors of prices or price movements. It seems obvious that direct estimates derived from prices collected by price collectors will be of high quality in terms of concepts, such as quality adjustment, and properly checked for unusual changes or replacement items. However, this collection provides relatively small quantities of data (though these may still form large datasets), and therefore, estimates derived from these data, particularly when disaggregated by regions or some similar small domains, are expected to have large variances. Scanner and/or web scraped data by contrast are not of such high quality but are probably good predictors of price change. But there are likely to be other variables which are also important, such as the type of outlet, its turnover (since for example small turnovers require higher mark-ups to make businesses viable), and possibly other variables. It is a challenge to assemble datasets which contain all of these variables, and we do not know of any examples where this kind of modelling has been attempted. Therefore, we also await information on the predictive power of these models.

The second challenge is, that the modelling will necessarily apply an element of smoothing to the price information which is the target of the CPI measure. If we are prepared to postulate an underlying smooth change in prices, and to say that this latent variable is the target that the CPI should measure, then this might be a reasonable approach. It is however also important to consider, whether prices might move less smoothly, particularly at times when there is a sudden economic downturn and things are not following their usual relationships (Smith and Lorenc, ress), and therefore whether this kind of modelling approach would generate the required results under all the circumstances required by an official measure of inflation.

An alternative use of models is to use the different information in the strcutured price collections and the alternative data sources in times series models which can predict or nowcast estimates of inflation using the correlations between the series. Powell et al. (2018) use daily web-scraped prices (for a restricted range of goods) to indicate short-term movements in prices, and examine the relationship between these and the official (sub)index for these goods, and show that it is possible to predict the official index using this kind of information. This type of modelling could be extended to exploit correlations between the prices of different types of goods sold in the same types of outlets.

It is again not particularly difficult to see how these models could be extended to regional price indices, either simply by considering each region in turn, or possibly through a multivariate model, since the number of NUTS1 regions in any country is generally quite small. These approaches require further research.

2.a.3.2.4. Price and quantity data

Lastly, there has been a lot of interest in using scanner for price indices because it provides price *and* quantity information. This potentially allows the elementary aggregate indices, used, because there is no quantity information, to be replaced by weighted versions which would be expected to more accurately reflect the different contributions of different goods. This should improve the relevance of the calculated index. For the time being however, scanner data have only partial coverage of the population (of transactions), and the ability of a National Statistical Institute to process these large quantities of data can reduce the availability further. Zhang (2018) suggests that one solution may be for the retailer do process these data and send calculated statistics (according to a defined methodology) to the statistical office, which would combine them into an aggregate index.

A second, but smaller challenge with scanner data is that the quantities are not associated with household variables, so no breakdown is available to produce bespoke weighting for particular groups of the population (for example, pensioners). Indeed, it is not even known whether it is the household sector which is making purchases. For many products, we may be happy to assume that the contribution of other sectors is negligible, but in some products this assumption may not be tenable.

Nevertheless, if the challenges can be worked out for a national index, the disaggregation to local indices should be relatively straightforward. The scanner data should be coded by store, and therefore regional datasets can be constructed and used in the calculation of regional CPIs.

2.a.4. Experimental UK regional consumer price inflation with model-based expenditure weights

2.a.4.1. Introduction

For a long time, users of price statistics have suggested that regional indices of consumer prices would be valuable in understanding how inflation varies across the United Kingdom (UK), and whether there are important differences in regional inflation (Department of Employment, 1971, Fenwick and O'Donoghue, 2003, ONS, 2013) para 3.13). The official position has been that the number of price quotes is too small at a regional level to support the calculation of indices, and it has not been a sufficiently high priority to invest in additional price collection for this purpose. Some limited information from the Office for National Statistics (ONS) on variation in regional prices has been made available through publications on Relative Regional Consumer Price levels (RRCPLs) (Baran and O'Donoghue, 2002, Wingfield et al., 2005, ONS, 2011), which have used information from additional price collections made every six years to adjust Purchasing Power Parity (PPP) statistics. PPP prices are collected in the capital city, and a periodic exercise is undertaken to adjust indices to represent the whole country. RRCPLs are spatial price indices which show the differences in price levels between regions (relative to a reference region = 100), but are not temporal indices designed to show price change (inflation, relative to a reference time = 100). Because of the methodology and differences in the weights on each occasion, RRCPLs are not satisfactory even for a once every six years approximation of regional inflation. The focus of this article is on a regional index that can measure the temporal differences, (inflation), rather than the spatial differences.

The Consumer Prices Index (CPI) is used as a national measure of inflation in many countries. Some countries have also attempted the development of inflation measures at a sub-national level. In the US, the construction of the CPI includes a regional aggregation phase, and therefore component indices at the regional level are part of the standard outputs (Bureau of Labor Statistics, 2018); Japan also produces regional indices (e.g. Statistics Bureau of Japan (2020); see also Nagayasu (2011)). Weber and Beck (2005) compiled a database of regional prices for Europe where they found regional or city indices in Austria, Finland, Germany, Spain and Portugal. In Germany, the Federal Statistical Office publishes regional CPIs for the 16 federal states (Statistisches Bundesamt, 2020) and there are also publications with regional indices for 401 German districts relying on an econometric model (Kosfeld et al., 2009). The national German weighting pattern is updated every five years using the "Einkommens und Verbraucherstichprobe" (Statistisches Bundesamt, 2018), but the sample size is insufficient for the estimation of regional weights. Other countries with regional inflation measures include Poland (Gajewski, 2017), Russia (Brown et al., 2018), Indonesia (Purwono et al., 2020), South Korea (Tillmann, 2013) and Turkey (Yesilyurt and Elhorst, 2014, Duran, 2016).

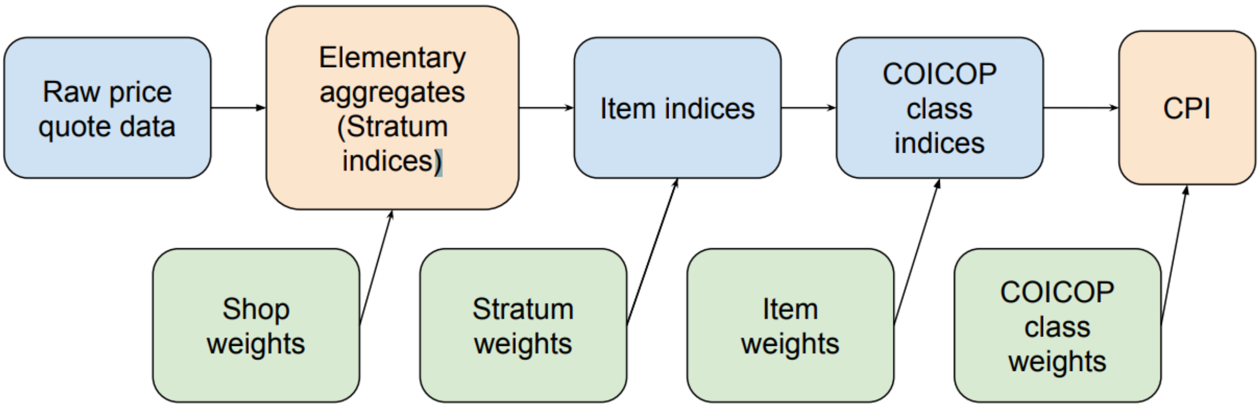
Other countries, including the UK, do not have a regional aggregation phase, so a special exercise is needed to produce regional CPIs. There has been interest in regional indices over many years in the UK. The Chancellor of the Exchequer announced work on regional prices in 2003 Fenwick and O'Donoghue (2003), which was translated into the development of RRCPLs. Although these have been published each time PPP data collections have taken place, there has not been any substantial development of these statistics. Fenwick and O'Donoghue (2003) also discuss the potential for regional temporal indices, but conclude that this needs further development; the annex to their paper lists the issues which make such a development challenging.

Economists have an interest in regional variations in price inflation (and more widely in regional differences in the cost of living, which is not so easily defined or calculated). Borooah et al. (1996), Hayes (2005) and Rienzo (2017) have all attempted to calculate regional versions of a consumer price measure for the UK with simplified methodology and based on available data sources. Regional variations in price levels measured by PPPs are also important inputs in local economic analysis. For example, Marchetti and Secondi (2017) estimate regional household consumption expenditure adjusted for differences in regional PPP in Italy, and Marchetti et al. (2019) use regional (province)-specific prices to adjust the national poverty threshold in Italy. Both of these applications use small area modelling approaches to make predictions of the variables of interest at local levels.

We assess the feasibility of producing a regional CPI measure for comparison of inflation rates between regions of the UK, that is, regional temporal indices, based on the official data collections. The UK has twelve statistical regions, including Wales, Scotland, Northern Ireland and nine regions of England; these are the Nomenclature of Territorial Units for Statistics (NUTS) 1 statistical regions. This chapter aims to develop a regional CPI rather than a CPIH (CPI including owner occupiers' housing costs), because of the additional complications caused by these additional housing costs which require regional measures of imputed rent and council tax data. Although the ONS prefers the use of CPIH to measure inflation, for simplicity we focus on the CPI.

At the national level, the UK's CPI is based on an extensive price collection. Elementary aggregates are calculated as unweighted geometric means of the relative prices within each stratum. The stratum level is based on variations by region and/or shop type (independent vs multiple). These geometric means can be thought of as stratum indices. The weighted arithmetic mean of the stratum indices using the sampling weights gives the item indices. For further aggregation, item indices are weighted together in proportion to the national consumption of the item, derived from national accounts estimates of expenditure, Living Costs and Food (LCF) survey data, market research data and other sources, including administrative data. The ONS uses the LCF data from year $t-2$ for the expenditure weights of the CPI in year t, giving a Lowe index. Weighted arithmetic means are then calculated at the class level based on the Classification of Individual Consumption by Purpose (COICOP) classification framework (more details in the next section). Finally, these class indices are aggregated using class expenditure weights to produce the national CPI. For further details of the CPI calculation, see ONS (2019) and also a simplified flow diagram is shown in Figure 2.a.1.

Figure 2.a.1.: Simplified flow diagram of the calculation of the UK national CPI.



The derivation of the CPI requires data sources to determine the price changes of certain goods, and also sources for the expenditure weights to determine both the basket composition and the amount of money spent on the goods. All these data sources are collected to ensure that the national CPI can be used to calculate reliable inflation rates, with suitable sampling designs and sample sizes to ensure that it is nationally representative. For the development of a regional CPI, these data for prices and expenditure are partitioned into the separate regions. The reliability of a CPI at the regional level will be compromised, because the reduced sample size leads to reduced precision of the estimates. We consider this the primary limitation to the development of reliable and temporally stable regional CPIs.

A second limitation is the accessibility of regional-level data sources. As mentioned, at the national level expenditure weights are calculated using many data sources. However, not all of these sources are readily available for each region. For example, at the time of researching, there was no regional equivalent for the National Accounts, though since then ONS (2018a) has published experimental household final consumption expenditure (HFCE) at the regional level. For future research, this regional HCFE data could be investigated to give balanced regional expenditure weights. Other data sources, such as administrative data, are not very accessible, but the LCF survey data is accessible and also has region identifiers, meaning that expenditure data from the LCF survey can be used to estimate regional expenditure. For this reason, we use only LCF survey data to estimate the expenditure weights. Price data do not have the same issues since the data are readily available with region identifiers, and are used directly.

The aims of this research are first to assess the feasibility of calculating an experimental regional CPI series using accessible data sources. Second, we investigate model-based methods to overcome the primary limitation of the reduced sample sizes. We look at smoothing and small area estimation (SAE) methods that may provide a means to improve the reliability of regional CPIs without having to collect additional data.

The structure of this chapter is as follows. In section 2.a.4.2, we provide more background on the conceptual framework of a regional CPI as well as background on available data sources and the COICOP classification. In section 2.a.4.3, we present methods for constructing an experimental regional CPI with just the LCF survey data and publically available price data. We also assess the experimental regional CPI series for 2010–2016. In section 2.a.4.4 we investigate the use of smoothing and small area estimation approaches to estimate the regional weights, intending to improve the regional CPI series. Finally, in section 2.a.4.5 we discuss the results and suggest further research.

2.a.4.2. Structure, data sources and COICOP classification

2.a.4.2.1. A conceptual framework for regional CPIs

The conceptual framework is set out in chapter 2.a.2.

2.a.4.2.2. Data sources

To develop an experimental regional CPI we use price quote data for the prices and LCF survey data for the expenditure. Monthly price quote and item index data for the UK are available from the ONS website from January 2010 (ONS, 2020b). The price quote data provides prices for items (goods or services) with corresponding information about the shop type, region, validity and stratum weight. Not all prices are represented in the price quote data, because many have nationally defined pricing and are collected centrally – approximately 45% of the weight of the basket is comprised of these centrally collected items. These central items are reported in the item index data sets, which report the indices and weights for the national CPI, but do not include a regional breakdown. Almost all items that contribute towards the CPI are reported, except 1-2% of the basket weight that is not represented due to disclosure control. For the regional CPI, the price quote data will need to be partitioned by region to calculate the item indices for each region, and then national level indices used for those collected centrally.

LCF survey data were obtained for the years 2008-2014 (Department for Environment, Food and Rural Affairs and Office for National Statistics, 2020) for estimating expenditure, which will contribute to regional CPIs for 2010-2016 (because the CPI is a Lowe index with expenditure data from an earlier period than the reference period). The LCF survey data included the household and individual-level data, both with region identifiers. The household sample sizes are shown in Table 2.a.2, and vary between regions and across years. Increases to the sample size in Northern Ireland (from 2016/7) and Scotland (from 2018) have more recently been implemented. The LCF survey data provides expenditure on products purchased by each sampled household. These products are classified according to the COICOP classification. An example of the COICOP classification is shown in Table 2.a.3, which also shows the labels for the different levels. The item level is also included, which is the lowest level of classification and sits below the COICOP hierarchy. Items are chosen by the ONS to be representative

Table 2.a.2.: Household sample size for LCF data by region, 2008 to 2014

Region	2008	2009	2010	2011	2012	2013	2014	Mean
North East	235	236	258	283	262	251	255	254.3
North West	592	582	596	647	623	585	588	601.9
Yorkshire and the Humber	491	484	485	521	521	462	459	489.0
East Midlands	405	393	413	455	425	424	440	422.1
West Midlands	469	527	470	526	513	526	470	500.1
East	532	499	515	543	563	497	498	521.0
London	472	464	476	536	490	480	407	475.0
South East	806	701	679	761	783	681	740	735.9
South West	502	518	495	507	493	429	468	487.4
Wales	265	272	261	251	266	246	222	254.7
Scotland	500	544	468	500	483	412	434	477.3
Northern Ireland	574	602	147	161	171	151	152	279.7

within a COICOP5 category and it is item prices that are reported in the price quote data. The LCF survey data reports expenditure at the COICOP-plus level.

We use the LCF survey data to estimate the mean (or equivalently the total) household expenditure by COICOP4 level in each of the twelve regions of the UK. For the national weights, the LCF survey data is one of multiple sources used to estimate expenditure. However, for the regional expenditure weights, we calculate directly from the LCF survey data without other sources, because they are not publicly available. The estimates of the mean household expenditure are then converted into relative weights (measured in ppt of expenditure).

It should be noted that because this article focuses on estimating expenditure weights from LCF survey data only, not all COICOP classes will be represented. Five classes were not reported in the LCF survey data in any of the available years. This is another reason why non-LCF survey data sources are used to construct the CPI. For any given household, there are usually many COICOP-plus categories which are not represented. This is expected, but it becomes problematic when only a few sample households in a region have recorded expenditure data for a particular category.

2.a.4.3. Constructing the experimental regional CPI

Ideally, the methods used to derive a regional CPI should be kept as close as possible to the national CPI. The conceptual framework for regional price indices is set out in chapter 2.a.2, and we have attempted to follow these concepts wherever possible. In adapting the methods of the national CPI to the regional level, we must take account of specific limitations, which fall broadly into two categories – those due to small sample sizes, and those due to only the LCF survey being used for expenditure weights. There are two immediate changes to consider for the regional CPI:

1. Remove regional information in stratum weights – Some of the strata in the national CPI are defined by the region. One of the purposes of the stratum weights is to adjust for the differing

Table 2.a.3.: Example COICOP classification for UK national CPI framework

Level	COICOP2	COICOP3	COICOP4	COICOP5	COICOP-plus	Item
Label	Division	Group	Class	Expenditure code	Category	Item
Example	Food and non-alcoholic beverages	Food	Bread and cereals	Bread	Buns, crispbread and biscuits	White sliced loaf branded 750g
Code	01	1.1	1.1.1	1.1.1.2	1.1.1.2.2	–
Number[a]	12	47	85	303	367	731

[a] Numbers tend to change over time, and the UK aggregates some categories too, so these should be taken as approximate

purchasing patterns across regions and shop types. However, when constructing a regional CPI, no adjustment for differences between regions is necessary, hence the stratum weights for regional indices reduce to shop type weights

2. No regional item weights – The item indices are to be weighted, but the LCF survey does not report expenditure at this level so regional item weights cannot be determined. This is a major shortfall in information for weighting the indices at the regional level. In the future, it will be important to develop methods to estimate these weights to refine the calculation of regional indices.

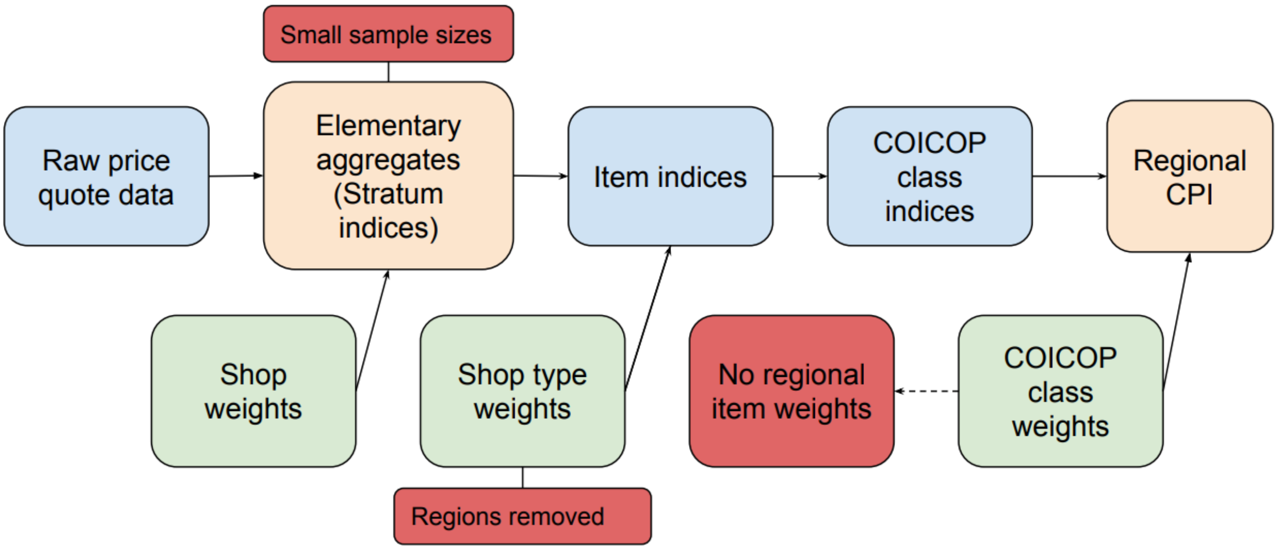
Another more general limitation is the aforementioned issue of smaller sample sizes at the region level, for both price and expenditure data sources. These three limitations lead to an adjusted regional CPI framework shown in Figure 2.a.2.

Although it is clear that surmounting these limitations requires us to approximate the target concept and therefore renders the regional CPI less reliable than the national CPI, the extent of this unreliability remains to be determined. We assess this by constructing an experimental regional CPI.

2.a.4.3.1. Regional CPI price aggregation

The steps taken to construct the provisional regional CPI, with commentary on the issues which arise, are as follows. The first step is to collate the price quote and item indices data. Of the 713 items with non-zero weights listed in the 2016 item indices dataset, 548 items were available in the price quote dataset, so regional versions could be calculated. This leaves 165 items (23.1%) which have only national series, and which will therefore not contribute to differences in prices between regions (though they may have different effects in different regions if they are weighted differently). For 2016, the 713 items account for 98.3% of the total weight. Therefore, we exclude items representing approximately 1.7% of the weight, and rescale the weights accordingly. For a more formal regional CPI, these missing

Figure 2.a.2.: Flow diagram for calculation of the regional CPI.



items would need to be included. The 548 items in the price quote dataset made up 53.9% of the weight. Similar percentages were observed in the other years prior to 2016.

Next, from the price quote data, elementary aggregates are calculated within each stratum, as the weighted geometric means of the price relatives. We use the geometric mean (Jevons elementary aggregate formula) for all items, though in the national CPI the ratio of averages (Dutot formula) is sometimes used (ONS 2019). All items have shop weights, which are used to get the weighted geometric mean of all items in each stratum; these are the stratum indices. Table 2.a.4 shows the frequencies of the number of prices used to calculate the stratum indices in each region.

There are large numbers of strata which have extremely small sample sizes, including 5.4% with only one price measurement. The majority of the strata with small sample sizes are for independent shop prices. Like all estimates, the geometric mean of very small samples is not very reliable, highlighting the primary limitation of constructing the regional CPI. Although it is inadvisable, for this experimental regional CPI the small sample sizes are treated as if they are satisfactory and the regional elementary aggregates are used. With the prices aggregated into the stratum indices, the item indices can then be calculated by taking the arithmetic mean of these elementary aggregates weighted with the shop type weights. As mentioned, the shop type weights are merely the stratum weights with the regional weights ignored. Once the item indices are calculated for each region the available national item indices not represented in the price quote data can then be added to give the full set of item indices required to calculate COICOP class indices (except for the 1.7% excluded).

The next step is to aggregate item indices to give regional COICOP class indices. However, as mentioned, the item weights cannot be determined at a regional level so this cannot be calculated using the available data sources. To overcome this problem we use the national (subscript n) item weights w_{nk} and class weights w_{nc} to first get national proportions of item k within class c. We then

Table 2.a.4.: Numbers of strata with $1, 2, \ldots, \geq 10$ price quotes for each region in 2016

Region	1	2	3	4	5	6	7	8	9	10+
North East	600	584	415	488	803	1,098	1,133	323	204	2,867
North West	381	497	413	251	151	128	151	169	226	6,560
Yorkshire and the Humber	381	397	303	209	264	351	387	438	502	5,714
East Midlands	422	247	164	127	226	361	525	717	794	4,929
West Midlands	307	357	373	375	338	446	324	399	618	5,235
East	616	200	327	449	209	172	260	308	378	5,979
London	219	97	192	275	302	336	465	492	469	6,365
South East	356	523	269	273	136	142	184	184	168	6,931
South West	316	220	316	444	460	293	349	367	442	5,707
Wales	838	247	328	397	723	1,251	1,001	400	193	2,917
Scotland	707	387	219	230	222	153	221	225	317	5,654
Northern Ireland	580	919	784	1,031	1,435	592	366	257	368	2,230
Total	5,723	4,675	4,103	4,549	5,269	5,323	5,366	4,279	4,679	61,088
Total percentage	5.4%	4.5%	3.9%	4.3%	5.0%	5.1%	5.1%	4.1%	4.5%	58.1%

multiply these national proportions by the estimated regional (subscript r) class weight w_{cr}, which gives approximations of regional item weights $\hat{w}_{rk}$: $\hat{w}_{rk} = w_{cr}\frac{w_{nk}}{w_{nc}}$. This ensures that the item weights sum up to the class weight for each region. Note that w_{rc} can be estimated using LCF data, as described in the following section.

Finally, the regional item indices and weights are used to derive the COICOP class indices and also the unchained regional CPI for each region. This process was replicated for all available years where price quote data and LCF data were readily available. The regional CPI series were then chained together using the same approach as the national CPI (ONS, 2019) and indexed with January 2010 set to 100 for all regions.

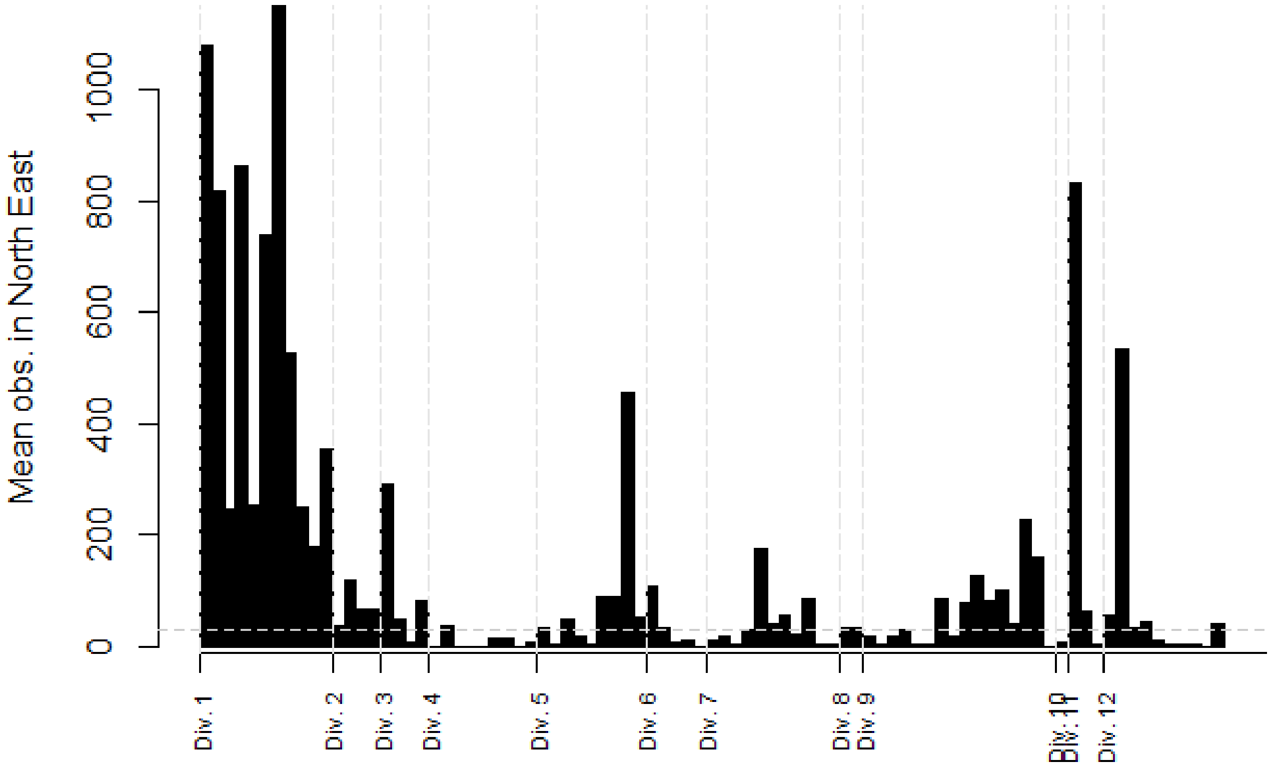
2.a.4.3.2. Regional CPI expenditure weight estimation

The LCF data provides expenditure for a sample of households in each of the twelve regions of the UK. To get expenditure weights for the regional CPI the expenditure should be aggregated within a certain COICOP level and region. We found that the COICOP class level (COICOP4) was the most suitable, as more detailed levels had too many zero expenditures and the COICOP group level was too general to give suitable specificity. Even at the class level, there were still problems due to some classes having few or no expenditures recorded at the region level. As such we first inspect how well each class was represented in the LCF survey.

First, we inspect the number of observations in each COICOP class. We use the term ‘observation’ specifically to mean any household with non-zero expenditure recorded from a given COICOP class. Observations of zero expenditure are still used for estimates, but for simplicity, we do not refer to these as observations. For the higher COICOP level of class, there can also be multiple observations (of more detailed expenditures) from the same household. Figure 2.a.3 shows the mean number of observations in the North East region across the available years. The North East is used as it generally has the fewest households sampled for the LCF survey. A reference line at 30 observations is included in the plot, which shows that there are many classes with fewer than 30 observations. At the COICOP class level, we can see that Division 01 ‘Food and non-alcoholic beverages’ is well represented, as well as classes 5.6.1 ‘Non-durable household goods’, 11.1.1 ‘Restaurants & cafes’ and 12.1.2 ‘Appliances and products for personal care’. On the other hand, 5 out of 85 classes (5.9%) have no representation in the LCF data for the available years, including water supply and sewerage, household repair services, hospital services and package holidays. These will all have zero weights in the provisional regional CPI. This is not just an issue for the North East region but for all regions, which are all reasonably comparable.

Although the number of observations is an important consideration, the actual expenditure must also be considered. The problem of having so few observations will be compounded when there is higher expenditure because it adds more weight to the regional CPI. For example, in the North East COICOP class 6.2.2 ‘Dental services’ and 12.4 ‘Social protection’ both receive fewer than 10 observations on average each year, but the former comprises 3.0 ppt of the total expenditure compared to 9.8 ppt for the latter. Since these estimates are to ultimately provide weights for the price quotes, the higher weights will have a greater impact on the resulting regional index. For this reason, estimates for class

Figure 2.a.3.: North East region: mean number of observations from 2008–2014 for each COICOP class with a reference line at 30 observations.



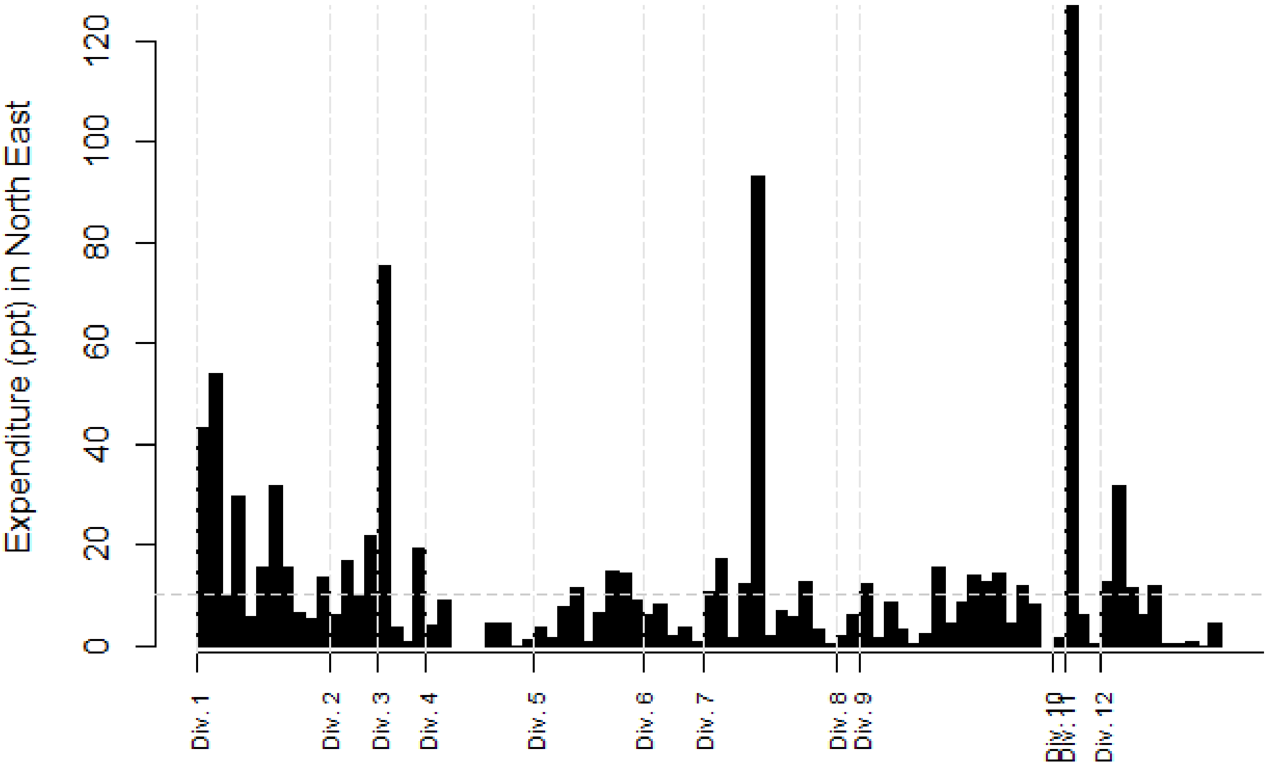
12.4 have greater potential to cause instability to the regional index compared to class 6.2.2. Hence we consider the expenditure estimates as well as the sample sizes for each COICOP category.

The North East mean expenditure in ppt for the COICOP classes is shown in Figure 2.a.4, with a reference line added at 10 ppt (1%) for comparison. There is a lot of variation between classes with five that make up more than 40 ppt and the remainder mostly under 10 ppt. The five largest classes, in order, are 11.1.1 'Restaurants and cafes', 7.2.2. 'Fuels and lubricants', 3.1.2. 'Garments', 1.1.2. 'Meat' and 1.1.1. 'Bread and cereals'. Note that the other regions also have highly variable relative expenditure across classes, and generally share the same largest classes.

To get direct estimates of the expenditure weights the following calculations are made. Let y_{ijct} denote the total household expenditure in pounds for COICOP class c in household j in region i and year t, and w_{ijct} be the provided household survey weight. Suppose that n_it is the household sample size for year t in region i, then the direct estimate of the mean household expenditure θ_{ict} can be estimated using:

$$\hat{\theta}_{ict}^{direct} = \frac{1}{n_{it}} \sum_{j=1}^{n_{it}} y_{ijct} w_{ijct} \tag{2.a.2}$$

Figure 2.a.4.: North East region: estimates of relative expenditure by COICOP class with a reference line at 10 ppt.



Then the relative weights in ppt can be calculated using:

$$\hat{w}_{ict}^{direct} = 1000 \frac{\hat{\theta}_{ict}^{direct}}{\sum_{\forall c} \hat{\theta}_{ict}^{direct}} \tag{2.a.3}$$

These weights can then be used to generate the regional CPI series for 2010-2016, which is shown for the direct estimates in Figure 2.a.10b. This series can be compared to the series calculated with the national weights (but regional prices) in Figure 2.a.10a. This comparison makes it clear that the differences between regions and the general volatility of the series come primarily from the expenditure weights rather than the prices. This is likely to be affected by small sample sizes with the consequent potential for influential weight estimates, but also the lack of the additional data usually used to adjust weights derived from the LCF data for use in CPI, which are likely to have a smoothing effect.

Due to the small sample sizes and lack of additional data, it is expected that there would be instability of the regional CPI series over time. To quantify this variability over time, we propose measuring the standard deviation of the first differences (SDFD) of the regional CPI series, that is $\frac{1}{T-2}\sum_{t=2}^{T}(y_t - y_{t-1})^2$. The higher this SDFD measurement, the more variable the monthly changes in the index. As a comparison, we find that the national CPI between 2010 and 2016 has a SDFD of 0.29. This provides an approximate guide for an appropriate level of temporal variability. We find that the SDFD for the series with regional prices and national weights in Figure 2.a.10a ranges from 0.36 for Wales, to 0.46 for Northern Ireland. This corresponds to 1.24 and 1.60 times that of the national level of variability, from using regional prices alone. The SDFD for the regional prices and weights from the LCF data ranges from 0.48 for North East, to 0.87 for Northern Ireland, corresponding to 1.65 and 2.99 times that of the national SDFD respectively. In all but two regions, the majority of the additional variability derives from the expenditure weights rather than the prices. These two regions were the North East and South West. Averaged over regions, approximately 60% of the additional variability according to the SDFD is due to the expenditure weights.

Due to the majority of the volatility of the regional CPI series coming from the expenditure weights, in the next section we explore methods to reduce this volatility. Specifically, we assess whether smoothing and SAE can improve this volatility.

2.a.4.4. Improving the expenditure weights

2.a.4.4.1. Smoothing methods

Since direct estimation of the expenditure weights leads to substantial increases in temporal variability, we assess whether smoothing methods can make improvements. There are many smoothing methods which can be used, but for simplicity, we test whether a three-year moving average of the expenditure weights substantially reduces the temporal instability. So instead of taking the mean estimate from observations in a given year, we include that year as well as the observations in the two preceding years (which approximates what would be available for real-time calculation of a regional CPI). This serves to increase the sample sizes, as well as strengthen the temporal correlation. Regional CPIs using these smoothed weights were calculated and the series are shown in Figure 2.a.10c.

Due to the removal of two years for the smoothing, these series cannot be set to 100 for 2010 like the full series. This reduces the comparability of these series; nevertheless, it appears the smoothing has decreased the variability as shown in Figure 2.a.10c. The SDFD ranges from 0.41 for North East to 0.72 for Northern Ireland. This amounts to an 8-20% reduction in the variability of the series. Hence there is evidence of improved stability compared to the one-year direct estimates, although only moderate improvement. This suggests that the three-year moving average approach does not eliminate all problems of volatility in the expenditure weights.

2.a.4.4.2. Small area estimation

Small sample sizes are demonstrably a substantial limitation to developing reliable regional CPIs in the UK. Like smoothing, SAE can potentially improve the reliability, as it utilises model-based methods to borrow strength from a wider or population-level data source to improve estimates for small domains. For a general overview of SAE methodology, Pfeffermann (2013), Rao and Molina (2015a) and Tzavidis. et al. (2018) are highly recommended. SAE is most effective when the wider data source has high quality, and when strong predictor variables are available within each region. In the case of the regional CPI, having region-level predictors of expenditure within COICOP classes will improve the precision of the estimates.

For the price quotes, it is difficult to utilise the beneficial aspects of SAE. This is because the sampling units are shops or prices. To effectively use SAE, comprehensive and informative data about the shop or price population would be required. This could include population numbers of shop types (independent and multiple) by region, number of supermarkets, and shop type by COICOP level. However, such data sources were not available so SAE could not be used. Furthermore, the expenditure weights rather than the price quotes cause more of the variability, hence SAE should be more effective at improving estimates of the weights.

SAE is much more feasible for the expenditure estimates. This is mainly because the sampling units for expenditure data are households, about which there is plenty of national and regional data available due to the population Census. The detailed national-level data for potential predictors (such as household types, total salary, total expenditure, head of household's age and the number of children) can be used in a model to get improved estimates of the expenditure for each class. We aim to use these small area models to give regional expenditure estimates with lower variances, and in particular look for them to be relatively smooth across years.

2.a.4.4.2.1. Fay-Herriot models

The SAE method used for expenditure estimation was the Fay-Herriot (FH) model (Fay and Herriot, 1979). The FH model is a commonly used region-level model for SAE (Rao and Molina, 2015a, section 4.2 & chapter 6), and comprises of two stages. The first stage simply models the sampling variation of the direct estimate which was defined in Equation 2.a.2: $\hat{\theta}_{ict}^{direct} = \theta_{ict} + \epsilon_{ict}$, where the sampling errors ϵ_{ict} are assumed to be independent and normally distributed with $\epsilon_{ict} \sim N\left(0, \sigma^2_{\epsilon_{ict}}\right)$. Out of the total household expenditure and the sampling weights we get regional direct estimators for each COICOP class and year. However, the variance of this estimator cannot be determined from only sample data. For this reason, the variance $\sigma^2_{\epsilon_{ict}}$ must be estimated. Possible options are the Poisson approximation or a bootstrap procedure. We used the bootstrap method of the `laeken` package (Alfons and Templ, 2013) in R (R Core Team, 2019a). For each COICOP class within every year and each region i, a bootstrap sample was drawn using the sample data with replacement. Out of the household weights and expenditures within each bootstrap sample, the corresponding direct estimator was obtained. The variance $\sigma^2_{\epsilon_{ict}}$ can be determined for all bootstrap samples. The second stage of the FH model is to fit a linear model which can be used to predict θ_{ict}: $\theta_{ict} = \boldsymbol{x}_{ic}^T\boldsymbol{\beta} + u_{ict}$ where $\boldsymbol{x}_{ic}^T$ denotes the region-level covariates, $\boldsymbol{\beta}$ denotes the regression parameter vector, and u_{ict} represents the random effects which, similarly to ϵ_{ict}, are assumed to be $u_{ict} \sim N\left(0, \sigma^2_{u_ct}\right)$. Note that for $\boldsymbol{x}_{ic}^T$ there is no subscript for the year because the same covariates are used across all years for each class. The combination of the two stages of modelling leads to the combined FH model: $\hat{\theta}_{ict}^{direct} = \boldsymbol{x}_{ic}^T\boldsymbol{\beta} + u_{ict} + \epsilon_{ict}$. The estimates $\left(\hat{\boldsymbol{\beta}}, \hat{u}_{ict}, \hat{\sigma}\text{fl}^2_{u_{ct}}\right)$ of these unknown parameters can be estimated using a standard linear random-effects model. From this, the FH estimates can be derived as:

$$\begin{aligned}\hat{\theta}_{ict}^{FH} &= \boldsymbol{x}_{ic}^T\hat{\boldsymbol{\beta}} + \hat{u}_{ict} \\ &= \gamma_{ict}\hat{\theta}_{ict}^{direct} + \left(1 - \gamma_{ict}\right)\boldsymbol{x}_{ic}^T\hat{\boldsymbol{\beta}}\end{aligned}$$

where $\gamma_{ict} = \sigma^2_{u_{ct}}\left(\sigma^2_{u_{ct}} + \sigma^2_{\epsilon_{ict}}\right)^{-1}$. In cases when an area has zero observations the estimator simply becomes: $\hat{\theta}_{ict}^{FH} = \boldsymbol{x}_{ic}^T\hat{\boldsymbol{\beta}}$. Estimates of the precision of the FH estimates can be made using the mean squared error (MSE) which is estimated using restricted maximum likelihood (REML). Further details can be found in (Rao and Molina, 2015a, chapter 6). Once the estimates are calculated, the weights can be created using the same adjustment as in Equation 2.a.3. One of the strengths of SAE is borrowing strength from other areas. This is done by using the association between the region-level covariates and the regional expenditure. Region-level associations strengthen each region's estimate by using the region-level covariates, which are assumed to be informative. In the next section, we describe how the covariates were selected for use in the FH models. A challenge for the FH model in this situation is that there are only twelve regions. With so few regions, model assumptions are more difficult to assess, covariate associations are more likely to occur by chance and the number of covariates that can be included in the model is restricted because there are few degrees of freedom. Another notable limitation in the application of the FH model to the expenditure weights is that some classes have no expenditure in the region. This becomes problematic, as zero expenditure for more than a few regions will lead to violations of the normality assumptions. For some classes, a zero-inflated model (Pfeffermann et al., 2008, Chandra and Sud, 2012) may be beneficial, which we leave for future research. The same FH model was also applied to the three-year averaged data to assess the collective impact of both smoothing and SAE on the regional CPI series.

2.a.4.4.2.2. Covariate variable selection

The LCF survey provides a large number of variables at the regional level which can be used to estimate expenditure. These variables relate to socioeconomic status, household composition and household features, e.g. tenure type, number of adults, weekly income. These variables were aggregated to the region level. For a FH model to be effective at estimating expenditure, the region-level covariates should be predictive of the expenditure of the COICOP classes. The challenge is to select the best combination of variables that ensures the relationship is predictive but not over-fitted to the sampled data. This over-fitting is especially a concern since there are only twelve regions, and hence twelve points from which to fit a model. Furthermore, the covariates should not have high multicollinearity, as this can greatly exaggerate over-fitting. Over-fitting will lead to small area estimates with under-estimated precision, as well as overly biased point estimates. Hence the explanatory variables must be selected carefully.

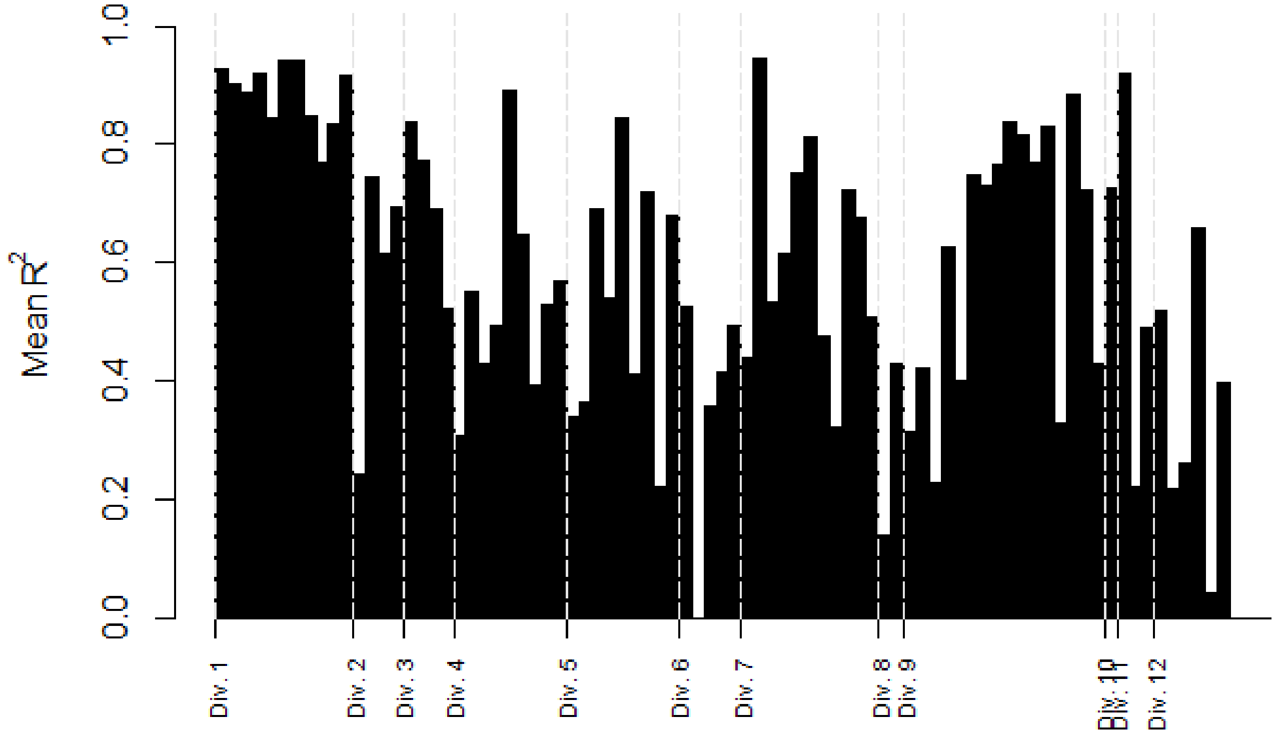
The variables were chosen based on associations with the class expenditure for a pooled data set across all years from 2010 to 2016. This ensures that the covariate variable is predictive across all years rather than a certain year. This will also ensure consistency across time. A forward selection approach using AIC was used to select the variables for each COICOP class, with at most five selected variables. We made five the maximum since any more would be superfluous when estimating only twelve regions. At each step, the multicollinearity was assessed using the variance inflation factor (VIF). If the VIF was greater than ten then no more variables were added. This ensured that a minimal number of variables was selected and that none of the variables was highly collinear.

2.a.4.4.2.3. Model assessment

A FH model relies on a strong level of prediction with explanatory variables. Figure 2.a.5 shows the R^2 values of the fitted linear models averaged over the years for each COICOP class. It shows that in Divisions 01 and 02, which includes food and alcoholic beverages, the R^2 is generally high, but some classes have low R^2 values. For example, COICOP class 12.6.2 'Other financial services' with a mean

R^2 of just 0.03, which will be unlikely to improve the expenditure estimate in the FH model.

Figure 2.a.5.: Mean R^2 over 2008-2014 for the chosen models for each COICOP class.



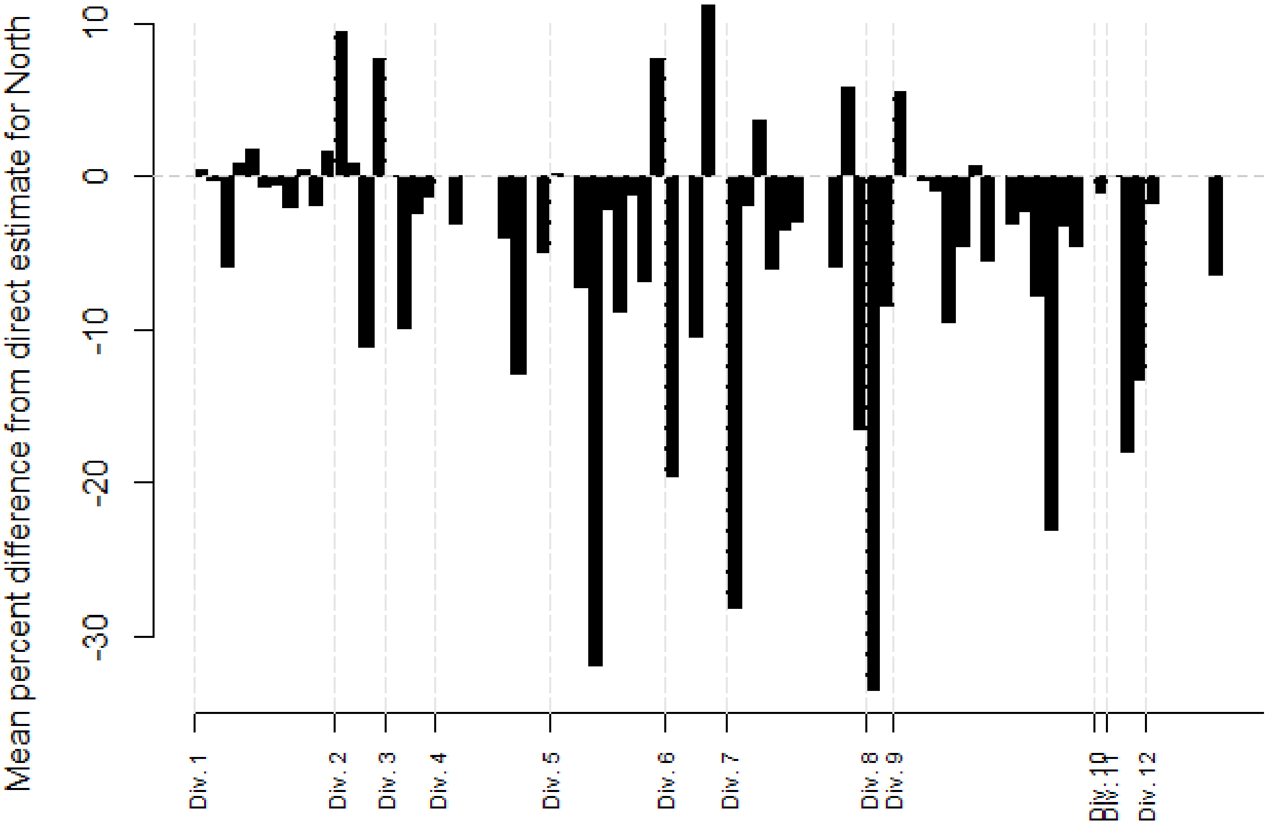
With the FH models fitted, it remains to be seen what the effect on the stability of the expenditure estimates is when we use the model predictions in place of the direct estimates.

As part of the model assessment, the assumptions of the models were checked, particularly the normality assumption of $u_{ict} = x_{ic}^T\beta - \theta_{ict}^{FH}$. A Shapiro-Wilk test was used, and based on this test, there was evidence to reject the normal distribution for many classes for each year. Across years, between 26 and 43 (out of 80) classes were not rejected, with a median of 36. However, as the focus is on improving the temporal stability of the estimates, we include all the estimates in calculating the experimental index, even if there is strong evidence to reject the assumptions. Further development of the models may provide a better basis for such estimation.

2.a.4.4.2.4. Assessment of Fay-Herriot estimates

To assess the effect that FH estimation has on the expenditure, we first measure how different the FH estimates are from the direct estimates. Figure 2.a.6 again uses the North East region as an example, and shows the percent difference between the FH and direct estimate, averaged over the seven years. This reveals up to a 30% difference in the estimates with many COICOP classes showing non-trivial relative differences. Clearly, FH estimation has some effect, but it is unclear what effect this is. Some classes have no percent difference because FH estimates could not be calculated, where too many regions had zero reported expenditure. In total, FH estimates could not be calculated for 16 of the 85 COICOP classes (18.8%). In these cases, the direct estimates are used for the weights.

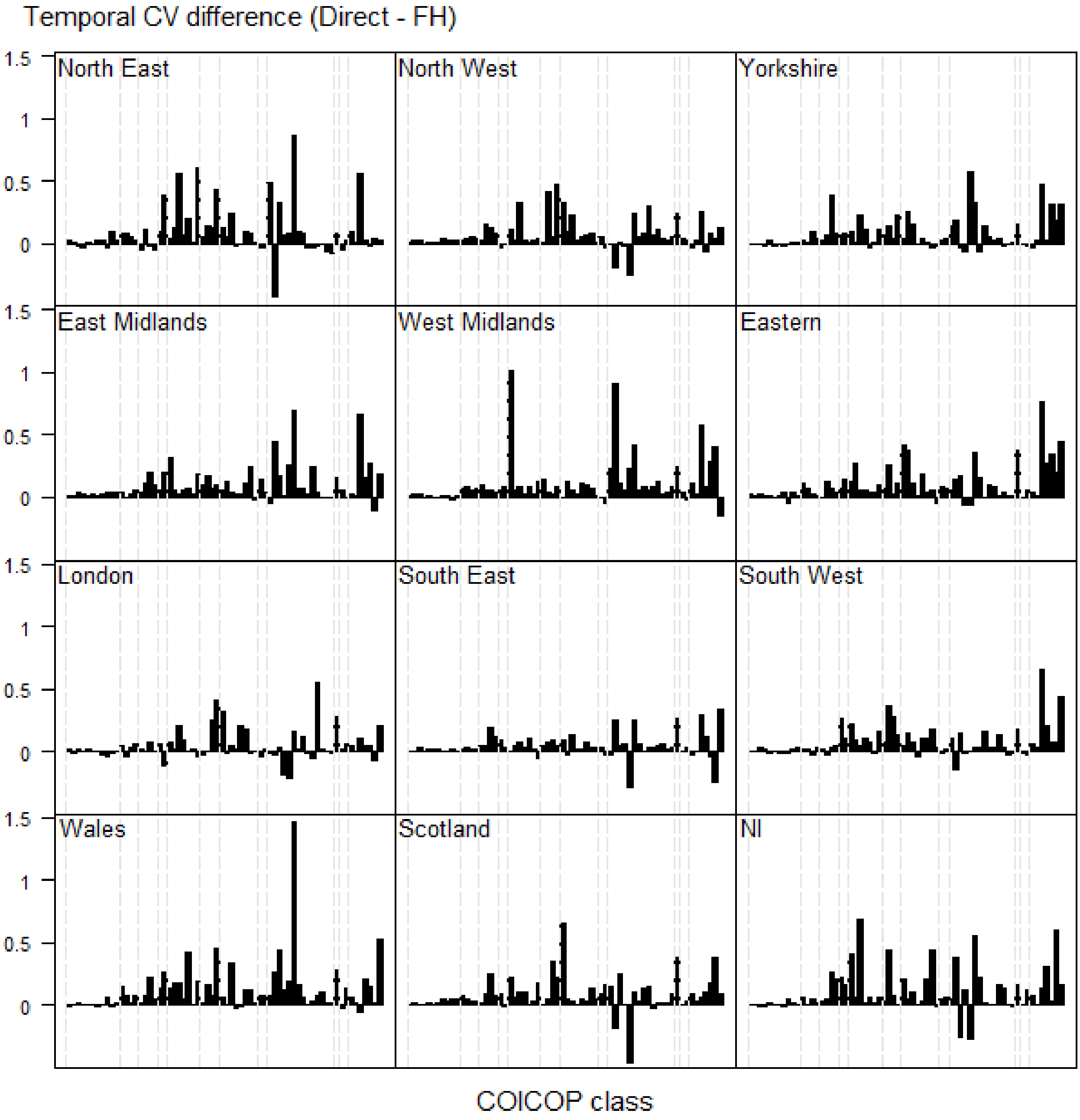
Figure 2.a.6.: North East region: Mean percent difference between FH estimate and direct estimate.



We expect that expenditure patterns are in reality rather stable between adjacent years, and change slowly as consumer spending is influenced by changes in products and their availability, particularly at higher levels of aggregation in COICOP. Therefore, we judge that the more stable the estimates are over 2008-2014, the better the estimates are and the more stable the regional CPI will be. This is because the instability is likely caused by small sample sizes. Note that some expenditure patterns may vary substantially over time, so an expert opinion on what level of variability is realistic would need to be considered too.

To measure this stability over time, we measure the variability of the yearly estimates of expenditure; this will include a small element of real change in expenditure patterns, but we expect that this is much smaller than the random variation we are trying to smooth using SAE. Typically the standard deviation or variance is used to measure variability, however this will not be appropriate in this case. This is because the variance is greater for COICOP classes with higher expenditure. To accommodate this, we use the coefficient of variation (CV) which is the standard deviation divided by the mean. This ensures the measure is standardised by the amount of expenditure, hence making the metric comparable across all COICOP classes. This 'temporal' CV is calculated using: $CV_{ic} = SD_t(\theta_{ict})/Mean_t(\theta_{ict})$ where SD_t and $Mean_t$ is the standard deviation and mean of the estimates across years t respectively. We use CV_{ic} to compare the stability of the FH estimates compared to the direct estimates. The lower this temporal CV, the more stable the estimates over time. Figure 2.a.7 compares this temporal CV for

Figure 2.a.7.: All regions: difference in temporal CV between direct and FH estimates.



the direct and FH estimates for all twelve regions. A positive difference in temporal CV indicates that the direct estimate is less stable, which is the case for the vast majority of classes and regions. This suggests that FH estimation is generally improving the stability compared to the direct estimates. Notably, in well-represented classes like in Division 01 the differences are very small.

Figure 2.a.8 displays the North East region estimates of γ_{ict} for each class averaged over all seven years. The higher the value of γ_{ict} the more the FH estimate utilises the data directly as opposed to the model-based component. There is a lot of variation between COICOP classes, ranging from 0 to

0.7, so there is no clear trend about what types of classes have higher values of γ_{ict}. Again, regions other than the North East showed similar patterns.

Figure 2.a.8.: North East region: mean estimates of γ_{ict} across years for each COICOP class.

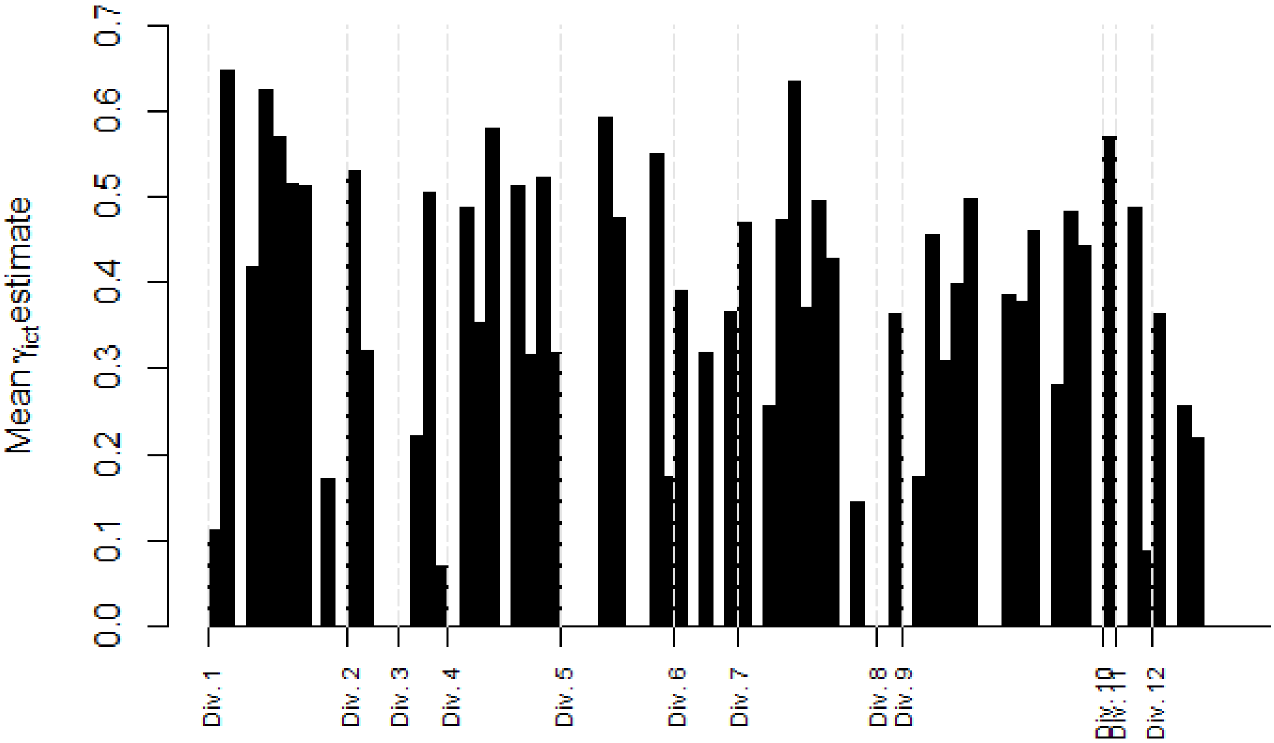


Table 2.a.5 reports the ten COICOP classes that have the greatest improvements in temporal CV due to FH estimation. These ten classes have generally few observations, ranging from 6 to 61. Interestingly, the R^2 values are not particularly high, which suggests that FH estimation does not require strongly predictive explanatory variables to provide additional stability to the estimates. Classes 6.1.2/3 'Other medical and therapeutic equipment' and 9.2.1/2 'Major durables for in/outdoor recreation' each have relatively large ppt values, which shows that it is not just the trivially small classes (such as 12.7 'Other services') that show improvements. The last column in Table 2.a.5 shows the mean values of γ_{ict} which range between 0.10 and 0.38 which is quite typical of all classes.

Figure 2.a.9 shows more broadly the effect of sample size on improved stability due to FH estimation. A smoothing spline has also been added in red to give an idea of the average effect for varying numbers of observations. The results show that for COICOP classes with relatively few observations the benefit of the FH estimation is generally better although highly variable for all regions. We also see how the COICOP classes with many observations show negligible benefit from FH estimation. The improvement of FH estimation becomes reasonably small after approximately 100 household observations.

In combining all these results, we consider four attributes of the COICOP classes which relate to their suitability for SAE. These four attributes are:

Table 2.a.5.: Ten COICOP classes with the most improvement in stability due to FH estimation.

COICOP class	Temporal CV difference	Mean no. of observations	Mean ppt	Mean R^2	Mean γ_{ict}
9.2.1/2	0.46	6.5	9.11	0.23	0.15
12.3.1	0.37	58.2	5.64	0.22	0.10
10	0.23	11.2	5.23	0.33	0.16
5.1.1	0.23	61.0	3.01	0.39	0.26
6.2.2	0.23	20.6	6.23	0.53	0.32
12.7	0.23	72.1	0.46	0.40	0.13
5.3.1	0.22	39.4	4.40	0.34	0.33
7.1.1A	0.20	16.6	2.91	0.36	0.38
6.1.2/3	0.19	59.7	15.03	0.22	0.21
3.1.4	0.15	15.2	1.20	0.69	0.32

1. Observations recorded in all regions for all years – 63 out of 85 classes (74.1%) meet this criterion. This means at least one reported price within a class for each region and year.

2. The number of observations not being so large that SAE remains useful, chosen to be all COICOP classes where all regions have at most 100 household observations – 63 out of 85 classes (74.1%).

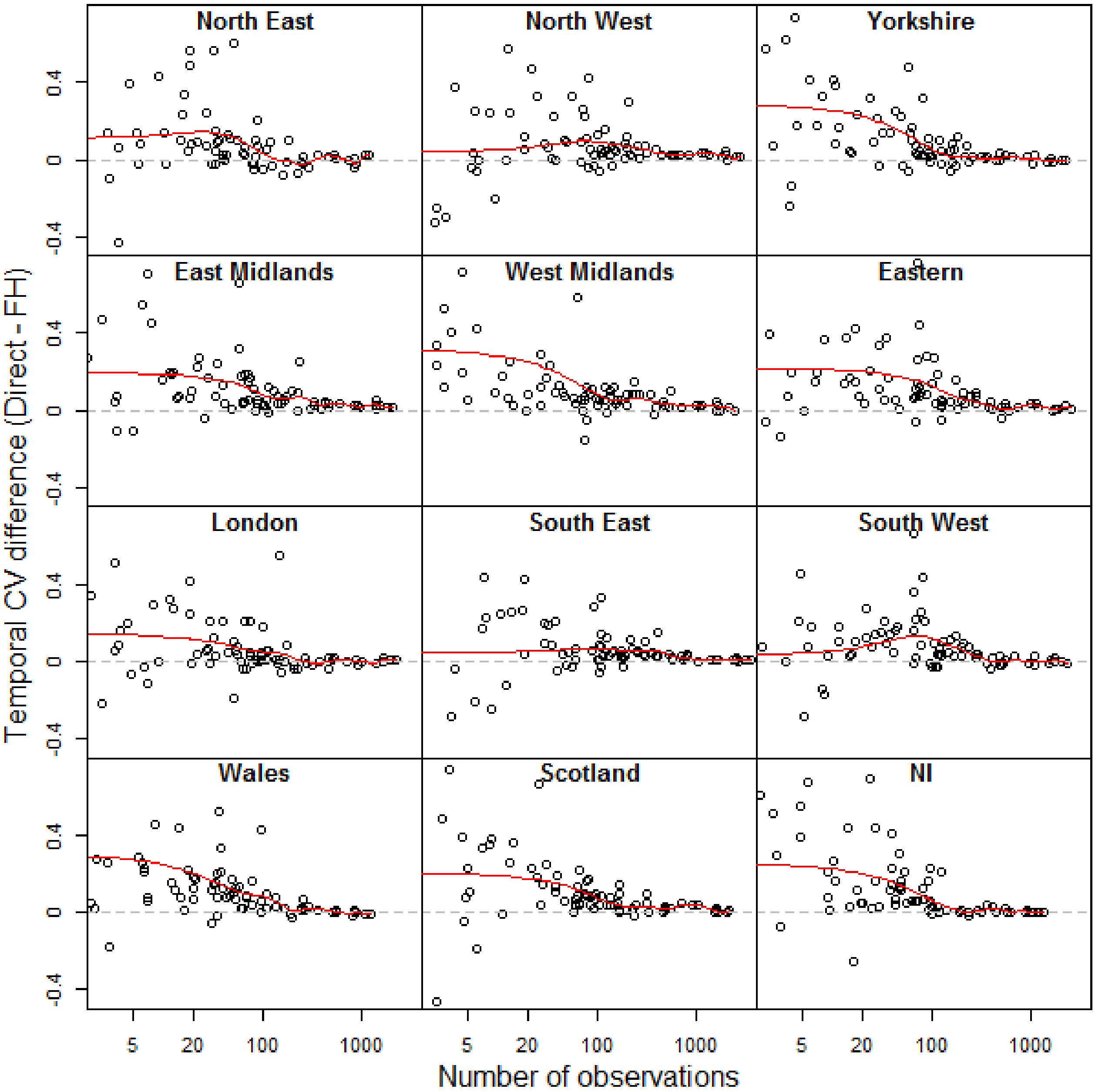
3. A non-negligible expenditure share in ppt, chosen to be the COICOP classes which have at least 0.5 ppt share in all regions and years (in line with the conceptual framework in section 2.1) – 58 out of 85 classes (68.2%).

4. A non-negligible (> 0.03) decrease in temporal CV for at least one region when using FH estimation – 60 out of 85 classes (70.6%).

These four attributes are not mutually exclusive, so the total number of COICOP classes which possess all four attributes is 36 out of 85 (42.4%). Hence, based on these criteria, 42.4% of COICOP classes have distinguishable improvements in the stability of the expenditure estimates through the use of SAE.

The regional CPI series with the FH estimates for each year and also the three-year average is shown in Figure 2.a.10d and Figure 2.a.10e. In comparison to the series with the direct weights, all regions but Northern Ireland appear much closer together. This is expected given that SAE adds national-level information, making the estimates closer to the national value. Generally, the observable differences between series with direct and FH expenditure estimates are not large. To assess in more detail we again calculate the SDFD for the two FH-based series, these are displayed for each region in Figure 11, as well as for the other regional CPI series. The SDFD metric shows that FH estimation does not generally decrease the variability over time compared to the direct estimates. There is a slight increase in the SDFD for most regions. So it appears that while SAE appears to make inter-regional differences smaller in the series, it does so without improving the temporal stability, and appears to

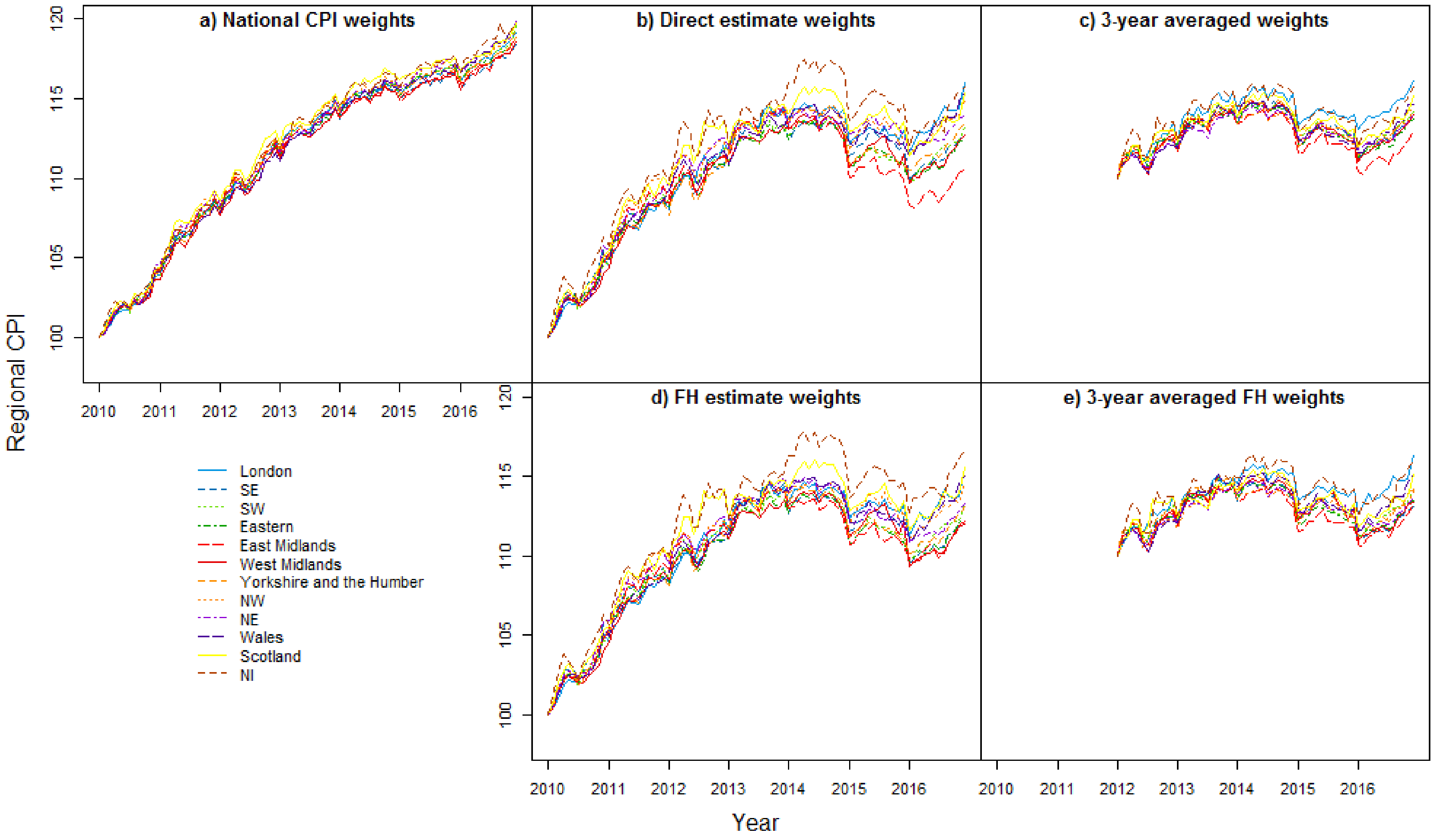
Figure 2.a.9.: Temporal CV difference by sample size within each region, with a trend line in red added. Each point is a COICOP class.



mildly increase variation over time.

The results show that smoothing and SAE using FH models for individual classes can improve the stability of the expenditure weights in some ways, but the smoothing appears to have a greater effect. The classes that benefit the most tended to have fewer than 100 observations, but also with enough observations for a good model to be fitted. Although FH models do ensure that the regional indices stay closer together, it does not provide reduced temporal variability for the regional CPIs.

Figure 2.a.10.: Regional CPI series for different expenditure weights. Note that 3-year averaged weights are rooted at 110 rather than 100.



2.a.4.5. Discussion of experimental regional CPIs

We have shown that it is possible to construct a regional CPI series from the available data sources. Although these experimental regional CPIs are somewhat useful, the reliability of specific components of the data and procedures is generally low. Small sample sizes create increased irregularities and variability over time. So, although it is feasible to construct regional CPIs, considerable development would be needed to ensure that they reliably represent the inflation within each of the regions. Importantly, we show that the source of this variability is generally from the expenditure weights rather than the price quotes.

Some further exploration of data sources would be possible. ONS has produced experimental regional HFCE estimates, which are balanced through the national accounts (ONS, 2018a). These are less detailed than the COICOP classification of the CPI, but they could be integrated into a framework of weight calculation. It would also be worthwhile exploring whether it is possible to calculate a version of the regional item weights which were not available for our analysis. Marchetti and Secondi (2017) go one step further by adjusting the regional expenditure estimates for differences in PPPs, which is potentially an important extension to the conceptual framework of chapter 2.a.2, though further investigation is needed to assess whether this would have a detectable impact on the regional CPIs.

Smoothing methods like the three-year moving average were shown to reduce the variability moderately. Although FH estimation improved temporal stability of the expenditure weights, there was no evidence that this reduced the temporal variability of the regional CPI series. We conclude that smoothing and SAE do generally make improvements to the series and the stability of the expenditure weight estimates, and that this can improve the reliability of the regional CPI. However, due to limitations in using only the LCF survey data and no other expenditure data sources, it is difficult to ascertain how well this would extend to a more formal regional CPI with additional sources. If similar irregularities remained present in a more formal regional CPI, then smoothing and SAE are plausible options. Fengki et al. (2020) have applied FH models to estimate regional CPIs in Indonesia (using the CPI as a target, because there are city CPIs available for modelling), with some improvement over direct estimation. But here too, further research is needed to deal with data deficiencies.

The FH model used has certain shortfalls which current methods could improve. These shortfalls are due to the model not considering the zero-inflated, longitudinal or compositional properties of the data. Methods have been proposed for small area estimation with zero-inflated data (Pfeffermann et al., 2008, Chandra and Sud, 2012), and it would be interesting to explore these for estimating expenditure weights. In particular, accounting properly for the observed zeroes in expenditure might allow modelling at a more detailed level in the COICOP classification.

An important extension to the FH small area model is to include temporal effects to account for correlation over time. Although this extension is not considered in the present paper, using it when working with panel data should help to improve the efficiency of small area estimates. Esteban et al. (2016) provide a comprehensive review of the literature on temporal extensions to the FH model that – among many variations – include a model with an autoregressive structure in the sampling errors (Choudhry and Rao, 1989) and a model with an autoregressive structure in the random effects

(Esteban et al., 2011). A further extension of the FH model is one that simultaneously accounts for spatial and temporal effects (Marhuenda et al., 2013).

A further line of investigation is to treat the estimation of weights as the estimate of a composition. Scealy and Welsh (2017) have developed an approach to estimate expenditure proportions as compositions within small domains of a population, different from the FH models, and Esteban et al. (ress) provide a FH model for compositions. Applying similar methods may give regional expenditure weights for the regional CPI with smaller variances and less change from year to year which would lead to fewer irregularities and smoother indices.

This chapter is intended to be a stepping-stone to the potential development of a regional CPI in the UK. Further research should be focussed on the adaptation of the proposed experimental regional CPI to a more formal one. Such an adaptation would incorporate the regional HFCE data, as well as all other sources that can be reduced to the regional level. If possible, regional estimates of owner occupiers' housing costs (OOH) and council tax could be developed making a regional CPIH which would be useful since it is the ONS's preferred inflation measure. Furthermore, if pursued more formally, it would be useful to have quality measures that a regional CPI or CPIH should aim to achieve. In particular, an estimate of the variance is needed, which is generally complex to calculate for consumer price indices; chapter 2.a.5 reviews the approaches to variance estimation in CPIs. Chapter 2.a.6 produces partial estimates for the variance of the UK national CPI, which could be extended to regional CPIs. This will help identify what is an acceptable level of temporal variability. Without this, one cannot definitively reach a point where a regional CPI can be shown to be reliable. Finally, a regional CPI should have regional estimates of the item level weights which cannot currently be calculated from the LCF alone.

2.a.5. A review of procedures for estimating errors in Consumer Price Indices, with special reference to sampling error

2.a.5.1. Introduction

The Consumer Price Index (CPI) is an important economic indicator with a wide use which encompasses macroeconomic policymaking to uprating of costs and benefits. It is a widely scrutinised statistic, not least because of the effect it has directly on the lives and budgets of individuals. Although the origins of the CPI depended on the foresight and interest of some key people, and on others for further development, the production of a national CPI has generally needed the resources of the state (O'Neill et al., 2017). In the UK, a CPI was introduced (under a different name) with the First World War, but it was not until after the Second World War that a well-designed system was instituted, including regular collections for prices, regular updating of the basket and regular surveys as a source of weights (Ralph et al., 2020, chapter 3).

Consumer price indices generally have a complex structure for data collection because of the different components, most of which have been collected through surveys (although recent opportunities to use scanner data and web-scraped price information may potentially change this). This has made it a significant challenge to produce estimates of sampling variances for CPIs, though the need for them has been acknowledged for a long time. The sampling structure for prices does not have common terminology in different countries. Here we will refer to the highest level of sampling as areas (in the US these are also described as PSUs and cities as they are the basis of city indices; in the UK they are locations; in Italy municipalities). The largest areas are included with certainty, and these are called self-representing areas. Within areas prices are collected from outlets (also called establishments, stores or shops). Commodities are classified (possibly in more than one hierarchical layer) into strata, called groups (item strata) from which representative items are selected (these are called entry level items the US). A price collector collects the price for a product (called an item in the US).

2.a.5.1.1. Historical development of variance estimation for price indexes

Jevons (1869) had already considered the error in deriving indices from the prices of a small number of commodities, and produced an estimate of the error in the change in his index of the value of gold (measured by the change in wholesale prices). Edgeworth (1888) continued this line of thought as part of a series of reports for the British Academy for the Advancement of Science, and he already concluded (p320) that one should "Take more care about the prices than the weights." Indeed his investigations were quite detailed, and it seems right to follow Balk (1987) and quote a little more detail of his conclusions (with Balk's highlighting in italics; the bold emphasis is Edgeworth's) by way of setting up the remainder of this review:

> "The error [of a price index] is found to depend *in a definite manner* upon **six** distinct circumstances. The erroneousness of the result is greater, the greater the inaccuracy of the data: the weights and the (comparative) prices. The erroneousness of the result is also greater, the greater *the inequality of the weights*, and the greater *the inequality of the price returns*. Lastly, the result is more accurate, the greater the number of the data and the smaller the number of omitted articles. These circumstances are not all equally operative. Other things being the same, *the inaccuracy of the price-returns affects the result more* than inaccuracy of the weights; and the *inequality of the price-returns more* than the inequality of the weights." (Edgeworth 1888, pp316-7)

Notwithstanding these early calculations of the variances of wholesale price indices, when the state took over collection and extended the range of price quotes available, the designs became considerably more complicated, and implementing a regular index calculation was the main priority. Therefore there was a period when there was little development in the production of quality measures for national price indices. Nonetheless it was not long before Bowley (1926, 1928) published further, considering the effects of correlation between price changes on the sampling error of an index, and the error components more widely. Bowley calculated sampling errors with autocorrelation on the index published by the *Statist*, for which all the prices (47 quotes) were published, and which provided a tractable model system. This approach does not seem to have been used on the Cost of Living Index produced by the Board of Trade in the UK, which was based on many more prices. Bowley also considered errors in index numbers more widely, and we return to this topic in section 2.a.5.4.

There was renewed interest in the calculation of sampling variances at least from Mudgett (1951, chapter 6) (see also Wilkerson (1967, section 2)). The first attempts at estimating components of variance due to specific parts of the sampling were in Sweden in 1953 and 1958 (see Dalén and Ohlsson

(1995)), and these followed work in the late 1940s on non-sampling error in the Swedish CPI by von Hofsten (see the re-analysis in McCarthy (1961, section V)). There were also forays into aspects of the variance of price indices by several authors — Banerjee (1956, 1959, 1960) and Adelman (1958).

However, the pressure to develop sampling error estimates was greatest in the USA, where Adelman (1958) evaluated the cost of a suitable system, and Kruskal and Telser (1960) and McCarthy (1961) were openly critical of the lack of quality measures. McCarthy (1961, section VII) developed an outline replication design which would allow variances and components of variance to be estimated, and this led to a redesign of the US CPI with the introduction of replicate samples from December 1963. These facilitated the construction of sampling variances which accounted for most of the components of sampling. The replicate sample areas are used for both price collection and the budget survey (Leaver and Valliant, 1995, p 554), which means that the effect of variance in both parts can be included in the same half sample estimates. This allowed estimates of variances to be made for changes in the index from 1964 onwards (Wilkerson, 1967), and the US was therefore the first country to produce good evidence for the overall sampling variability of its CPI. The Bureau of Labour Statistics (BLS) has generally continued this approach to calculating sampling variances since then, with some developments, particularly an update which added probability sampling for all the components of the CPI from 1978 (BLS, 2015, chapter 17) and some occasional experimentation to investigate the properties of alternative methods. As well as using the replicate samples for calculating the variance of the index, BLS has also used it to estimate variance components — which part of the sampling has the highest variance — as an aid to optimising the CPI sample (Baskin and Johnson (1995), extended using scanner data by Leaver and Larson (2003). A more detailed history of the progress of variance estimation work in the US is given in section 2.a.5.4 below, as it will be more comprehensible after the different variance estimators have been introduced.

The UK undertook some initial studies in support of a review of the possibility of producing regional price indices, where the effect of the sampling error might be more pronounced, initially in Department of Employment (1971, Appendix 1), where the effects of sampling errors in the prices and the weights were considered separately. This work was later extended by Fowler (1973), particularly with regard to the effect of sampling error in the weights. More details of the developments in the UK are given in section 2.a.6.2.

There was a general increase in interest in the measurement of the quality of CPIs during the later 1980s, with developments taking place in the Netherlands, Italy and Sweden. This led to several sessions at the biennial meetings of the International Statistical Institute. The Dutch undertook some work to estimate the variance of the CPI, mainly between 1985 and 1991, initially using balanced half sample methods (Balk and Kersten, 1986, Balk, 1989), and later moving to a design-based approach using Taylor linearisation (Balk, 1991). Unfortunately, the details of these latter derivations are unpublished.

At the same time Biggeri and Giommi (1987) undertook a review of strategies for estimating the sampling error of a CPI, coupled with some initial work on the variability due to estimating household expenditure weights in the Italian CPI. They suggested focusing on a sampling structure where the household budget survey (providing the weights) and the price collection were undertaken in the same sample areas, but were unable to match up areas at this level, and had to start with higher level aggregates. Later D'Alò et al. (2006) undertook further work in Italy to estimate the variance components (from sampling areas (municipalities), outlets, items and residual) for two groups of commodities; this allowed model variances for these groups to be produced, but they were not at that stage combined into an overall model variance.

The beginnings of work in Sweden also coincided with these developments elsewhere, with Andersson et al. (1987b) undertaking some preliminary studies on parts of the error. But this was followed by further development work to estimate sampling errors, where Dalén and Ohlsson (1995) used the cross-classified sampling of outlets and items as the basis for a design-based estimator of the variance attributable to price sampling. They claim this as the first calculation of sampling errors from a truly design-based estimator of the sampling variance, though Balk (1991) also seems to have taken a design-based approach. See the next section for more discussion of design-based and model-based approaches to variance estimation. Norberg (2004) evaluated this estimator in a simulation study and found that it was effective as long as there were indeed strong outlet and item effects, but that if these effects were weak the variance was overestimated. A similar procedure was used in Finland (Jacobsen (1997) in ILO et al. (2004, paragraph 5.99)). Dalén (1995) also considered other error components for the Swedish CPI in a complete framework.

In the UK further work was undertaken as part of a review of the Retail Prices Index in the 1990s, with unpublished reviews of variance component estimation in support of sample allocation in 1995, and of variance estimation in 1998; for more details see section 2.a.6.2.

In France, Ardilly and Guglielmetti (1992) derived variances accounting for the multiple stages of aggregation in prices in the estimation of the CPI, and based on a two-stage design for the price collection. As in most applications, they needed to make some simplifying assumptions about the sampling, and explored different approaches to imputing the variances where there were insufficient prices to estimate them directly. They commented that the weights might be important, but did not take account of the variation in the weights in variance estimates (except where they were temporarily zero because no prices were collected).

In Luxembourg some variances were calculated in 1998, based on a model which allowed for some correlation between price changes in the same establishment, and based on several simplifying assumptions about the sampling. See ILO et al. (2004, paragraphs 5.94-5.97) for a sketch of the approach, and Dalén and Muelteel (1998 in ILO et al. (2004)) for details.

In parallel with these developments in Europe, there was continued refinement of existing methods and development of alternative approaches in the US; these are documented in section 2.a.5.4.

In Italy the sample design was due to be made fully probabilistic from 2006, to support variance estimation. The proposed design contained some interesting features, with municipalities selected by balanced sampling and outlet samples for different item groups selected with positive coordination to maximise the outlet sample overlap (D'Alò et al., 2006). It is not clear whether this proposal was actually implemented.

The final entry in developments of variance estimates for a CPI belongs to Norway, where Zhang (2010) developed a model-based procedure, using similar estimators to Kott (1984) and Valliant (1991) for Laspeyres-type indicators, but also extending these to Paasche-type and, perhaps more interestingly, to elementary aggregate models corresponding to the Carli, Jevons and Dutot indices. The corresponding estimates are regularly calculated, but remain unpublished.

2.a.5.1.2. Overview

The main (published) developments in calculating variances and errors for consumer price indices have been made in rather few countries — USA, Sweden, Italy, Netherlands, France, Luxembourg, and UK. There are also applications of these approaches in Finland, an example from Brazil (Fava, 2007) which uses an estimator of the variance of the arithmetic mean of prices to approximate the variance of a geometric mean, and one from Iraq (Fatah and Ahmed, 2012), apparently based on the US approach. Several papers originated from India in the 1950s and 1960s and/or were published in Sankhyā, but this does not seem to have been translated into an estimate of the variance of the consumer price index in India.

Andersson et al. (1987b) say that "[t]he Swedish CPI has not been regarded, by its users, as an estimate of an unknown parameter. Rather, there seems to be a widespread agreement that the published value of the CPI is, by definition, the truth." This seems to be a general situation for CPIs, particularly since there is no single optimum choice of aggregation formula(e) to use. In most cases where error calculations have been made there are therefore challenges in explaining exactly what components of the quality and (for variances) what type of variance are being measured.

Some further work has been done on the sample size and allocation requirements for price indices, and this has often involved some estimation of the components of variance due to different parts of the design (e.g. Baskin and Johnson (1995), Leaver and Larson (2003), D'Alò et al. (2006)). This has sometimes resulted in estimates of sampling variances or components thereof as a by-product of the main objective.

The remainder of the paper consists of a discussion about design-based and model-based approaches to variance estimation for price indices in section 2.a.5.2. Section 2.a.5.3 gives an overview of the methodology for different approaches to calculating sampling errors (and most of these approaches can be applied to deal with variability in either the prices or the weights). Section 2.a.5.4 provides a more detailed history of the development of variance estimates for the CPI in the US, where much of the research into different approaches has taken place. Section 2.a.5.5 considers the range of types of errors which affect price indices, and the frameworks and approaches which have been suggested to combine them. Here, there is also consideration of how to combine variance estimates due to different elements of the CPI sampling processes. Section 2.a.5.6 gives a discussion of the effectiveness of different approaches, and highlights some areas where further research is needed to support quality measurement in national consumer price indices.

2.a.5.2. Design-based and model-based variance estimation

The tension between model-based and design-based approaches to thinking in survey statistics has lessened as it has been realised that some problems can only be approached through the use of models, but the different approaches have both been used in the calculation of errors in the CPI. The distinction affects, at least to some degree, how the resulting statistics can be interpreted. The strict definition of a (design-based) sampling error is the difference between the values obtained if the whole population could be sampled and the values obtained through a sampling process. Von Hofsten (1959) argues that the calculation of a sampling error for a price index is essentially impossible, because various parts of the sampling cannot be sensibly defined. This has not prevented multiple approaches to calculating such errors, and McCarthy (1961) makes a strong case for the existence and value of the sampling error concept for price indices. But there is certainly an element of truth that a range of assumptions and simplifications must be made to construct a suitable process for estimating the sampling error of a price index, and we document these in section 2.a.5.3. Von Hofsten's line of thinking leads naturally to the model-based approach, and we return to this below.

A design-based approach does not in fact need the whole population to be specified, but does need probabilistic selection at each stage, and the selection probabilities. The calculations do not require information on the distributions of the values (of the prices or the weights) in the population, though in practice long-tailed distributions containing outlying values may have an important effect on the estimates. These estimates have an interpretation in terms of the error arising from the sampling processes for prices and for the weighting information. Including both components together in an evaluation is challenging.

An intermediate approach is to seek to calculate the same design-based sampling error, but to use an approach with a model-based justification to calculate it. The replication methods in the BLS's CPI fall into this category, and Kott (1983) gives a model-based justification for why this approach produces satisfactory results. In some sense, variance estimates that work well under both model-based and design-based approaches are ideal, because they have natural interpretations in both contexts. Jackknife and bootstrap approaches are similar to the replication methods, and attempt to approximate the sampling distribution of the required statistic.

Several authors have espoused a model-based approach to estimating price indices, generally of the Laspeyres type (L-type, following the notation of Zhang (2010)). Kott (1984) seems to have been the first to take this approach, setting out a superpopulation model under pps sampling in a simplified unstratified design where the prices are assumed to be "nearly homogeneous", and using a modified (model-based) version of the HT estimator of the L-type index. He considers cases with and without autocorrelation in the price trends. Valliant and Miller (1989) and Valliant (1991) set up model-based estimators built on a simple autoregressive model for price change, the former for a one-stage design, the latter for a two-stage design with rotation. Both models admit a range of model-based estimators, and Valliant (1991) in particular identifies some such estimators which also have design-based interpretations and the corresponding variance estimators can therefore be regarded as approximations to the design-based estimators too.

Zhang (2010) chooses an explicitly model-based approach, and uses the residual variance to capture the extent to which the observed data vary around a specified model. Zhang's models are designed to be appropriate to the chosen form of the index, but do not have the property of going to zero as the sample size increases towards the population size. In this sense, the variances produced do not have the same interpretation as variances due to the sampling process, unless we make very specific assumptions about how the model also captures the sampling information. Some of the estimators in Kott (1984) are consistent, and some are not; but as Kott points out, (in the superpopulation framework) "consistency is an odd property to require of an estimator based on a sample of a specified size". Nevertheless, such an approach can be much more straightforward to calculate, and be a good indicator of the variability in the price index. Kott (1984) takes this approach further, and sets out the theory for a superpopulation approach to the design of a price index; adopting such a strategy would make the variance estimate and the design line up more clearly than in Zhang's approach where the design and variance estimation come from different paradigms.

Norberg (2004) undertook a simulation study involving both design-and model-based estimators of the variance. His model-based estimators are of the variance component type used by Baskin and Johnson (1995). He found that the model-based variance estimator generally worked well with synthetic data, but did not always produce results when based on simulations derived from real data, He concluded "[t]his estimator, however, is not practical for complex situations like this", and overall preferred one of the random groups estimators.

2.a.5.3. Methods for calculation of sampling errors

This section lists the methods that have been proposed for the calculation of sampling errors in consumer price indices, and discusses their strengths and weaknesses.

2.a.5.3.1. Design-based approaches with Taylor linearisation

Several authors have used standard results from sampling theory as a basis for deriving variance estimators accounting for the complex designs used in price index surveys. Andersson et al. (1987b) use this approach to assess the variance due to sampling outlets in the Swedish CPI; in their particular application sampling is πps and they make the (commonly used) simplifying assumption that sampling is with replacement. See their section 3.4 for detailed expressions. Ardilly and Guglielmetti (1992) also take a design-based approach to the variance of the French CPI.

As an example of the kinds of expression which result, in the US CPI Taylor linearisation was for some time used to give an approximation for the variance of the Laspeyres index between times s and t:

$$var(\hat{I}^{t,s}) \approx (\hat{I}^{t,s})^2 \left(\frac{var\left(\sum\limits_{i,m} \hat{w}_{im}\hat{I}^{s,0}_{im}\right)}{\left[\hat{w}_{im}\hat{I}^{s,0}_{im})\right]^2} + \frac{var\left(\sum\limits_{i,m} \hat{w}_{im}\hat{I}^{t,0}_{im}\right)}{\left[\hat{w}_{im}\hat{I}^{t,0}_{im})\right]^2} - 2\frac{cov\left(\sum\limits_{i,m} \hat{w}_{im}\hat{I}^{s,0}_{im}, \sum\limits_{i,m} \hat{w}_{im}\hat{I}^{t,0}_{im}\right)}{\left[\hat{w}_{im}\hat{I}^{s,0}_{im})\right]\left[\hat{w}_{im}\hat{I}^{t,0}_{im})\right]} \right) \tag{2.a.4}$$

where the $\hat{w}_{im}$ are weights and $\hat{I}^{t,0}$ are component indices for item stratum i and index area m (Leaver and Valliant, 1995). The variances and covariances were estimated with replicate samples (see section 2.a.5.3.2), but in principle these could also be derived analytically using Taylor expansions of ratios; nevertheless such a procedure becomes quite involved. There is a suggestion that Taylor linearisation may underestimate empirical variances for smaller sample sizes (see for example Andersson et al. (1987a)), and this underestimation and/or the errors of approximation, may add up in the different components of the formula.

Valliant (1991) also uses Taylor linearisation to derive variances, but in a superpopulation (model-based) framework. He makes the argument that these variances can also be considered as the design-based variances because the index formulae can also be derived from a two-stage cluster sample of outlets and prices within them. Valliant notes that for long-term price changes the number of covariances to be estimated can become very large.

Dalén and Ohlsson (1995) take the cross-classified design which arises from the independent selection of items and outlets and use this to derive a design-based estimator, again using Taylor linearisation to deal with the ratio form of the price index. Their equations (3.3)-(3.6) give the form of this estimator, which contains many terms and is not recapitulated here. They deal only with the variation in the prices, with further thinking about how the variation in the weights should be incorporated in Ohlsson (1995). Skinner (2015) extends the cross-classified sampling results, and in particular gives a bootstrap estimator (see section 2.a.5.3.3) which is computationally easier, though it has not yet been applied to price indices.

2.a.5.3.1.1. Strengths and weaknesses

Design-based approaches can be more efficient than replication approaches when the latter require the collapsing together of strata to make estimates, and the criticism that they need rederiving for different estimators is less pertinent for price indices where the form of the estimator does not change (Valliant, 1991). So, many of the standard criticisms of Taylor linearisation are muted in this case. The estimators may nevertheless be extremely complicated, and require several layers of approximation, whose aggregate effect is not clear without detailed investigation. Dalén and Ohlsson (1995) nevertheless produce a valid design-based estimator, and Norberg (2004) shows that this is effective when the cross-classification involved reflects strong effects in the data. There are real challenges in how to adapt the design-based approach to properly take account of imputation, and also how to deal with procedures for dealing with the transience of products, such as product replacement rules and adjustments for quality change.

2.a.5.3.2. Replication-based approaches

An approach which avoids the tedious calculation of variance estimators is to use replicate samples (also known as balanced half samples, balanced repeated replication). This is the basis of the longest-standing regular approach to calculation standard errors for a CPI, from the US (e.g. Wilkerson (1967), Leaver (1990)). Leaver et al. (1991) included the contributions of both the variance of the weights and the variance of the prices to the overall estimates, facilitated by having common sample areas for the two surveys (Leaver and Valliant, 1995, p 554). Replicates have also been suggested for individual elements of the variance including the weights (Balk and Kersten, 1986) and Koop (1986) outlines an approach to combining the variability of weights and prices even when the surveys take place in independently sampled areas.

Each sample stratum $h = 1, \ldots, H$ is divided into two parts, h_a and h_b called half samples. A series of replicates is constructed by choosing one of the half samples from each stratum, appropriately reweighted to give the correct population estimate, and using the variation among these replicates as a basis for estimating the variance. It is not necessary to use all $2H$ replicates, though in general the larger the number used the better the estimate obtained. The process can be made more efficient by use of a Hadamard matrix, a matrix containing 1 and -1 entries with orthogonal columns (Wolter, 2007), which makes the procedure balanced (in the replicates). The procedure is to take a column of the matrix, and to take h_a in stratum h if the h-th element of the column is 1 and h_b if it is -1. This set of half samples is used to make an estimate (of any statistic of interest, but in this case the index). Then the process is repeated with the next column; label the index estimated from the a-th replicate by I_a. If there are A replicates the variance estimator is then

$$var(I) = \frac{1}{A} \sum_{a=1}^{A} (I_a - \bar{I})^2 \tag{2.a.5}$$

where $\bar{I} = \frac{1}{A} \sum_{a=1}^{A} I_a$. Biggeri and Giommi (1987) give three further estimators in their outline of the method.

There are practical challenges in setting up a collection to operate this way. In the US the whole collection process for prices was replicated with selection of two sets of prices within self-representing metropolitan areas, or with paired selections of non self-representing areas. In particular a different sample of representative items was chosen in each replicate (by the same sampling procedure); the half-samples need not contain the full detail, so it is sufficient if the union of the samples provides sufficient detail for the calculation of the national index. In an ideal situation these replicate sample prices are collected independently, and then all of the variation in the collection procedures is accounted for in the replication variance. This is a singular advantage of this procedure, particularly with respect to price indices with their relatively complex data collection processes, that it can measure the variability due to non-probability but repeatable sampling procedures. For example there is no need to derive variances accounting for replacement indices or quality change as long as the procedures are used in the same way in both replicates.

If the replicate sampling can be repeated in further stages of the sample design, then the variances induced by sampling in the different stages can be estimated, and the information used to improve the efficiency of the sampling process. This was implemented in the US where a few cities had such additional replication, and this was used in decomposing the variance (Wilkerson, 1967), though with limited success because the variance components themselves have large variances. Later work on variance decomposition has relied on models, see section 2.a.5.3.4.

Outside the US where no other country uses a design with real replicates, the process is more usually approximated by dividing a single sample into two pieces, which therefore have a small negative dependence, which would not be present if the replicates were selected separately with replacement. If the sampling fractions are small, the difference from ignoring this dependence should be negligible, and this is the procedure used by Balk and Kersten (1986). In the US the procedure changed around 2012 from having two separately selected half samples to a single sample with prices randomly allocated to replicates (Shoemaker and Marsh, 2011); the results were substantially similar, though in general a little lower because the new method improved the balance between the two half-samples and eliminated one component of initial weight variation. It is however essential to ensure that price quotes are retained in the random group to which they are originally assigned. Redoing the randomisation period to period leads to substantial overestimates of the variance of changes relative to the original BHS method.

2.a.5.3.2.1. Strengths and weaknesses

The replicate samples approach has definite advantages in estimating the variance due to all the components of the sampling procedure, whether probability-based or not, as long as they are repeatable in the different replicates. Therefore variability due to imputation (Leaver and Larson, 2002), product replacement and quality adjustment are all included as long as the replicates are set up properly. There is a cost to running a system that really implements the replicates, but even in the US this has now been replaced by forming replicates from a single sample (Shoemaker and Marsh, 2011), and the same approach has been used in other countries (e.g. Balk and Kersten (1986)). And because there is no need to fit or assume a suitable model, there is no difficulty in asserting the objectivity of the resulting variance estimates.

If the sampling procedure generates small samples, replication can produce large variances. Shoemaker (2009) reported an unusual estimate of the variance of the US CPI which was traced to a single cell within the housing category, where a particular set of circumstance led to a small and heterogeneous sample. One of the two housing replicates as a consequence contributed approximately half of all the variance in the national CPI in this month. Shoemaker considered several alternative estimators, including jackknife and bootstrap estimators (see section 2.a.5.3.3), and while some were better in the affected month, they showed outlying variance estimates in other months which were not seen in the replicate samples approach.

2.a.5.3.3. Jackknife and bootstrap

Jackknife and bootstrap methods have become increasingly popular for variance estimation, and various developments have been made which enable them to be used in ever more complex designs. The jackknife is already in use in the US for special item categories which do not have replicates and therefore do not fit in the random groups method (BLS, 2015, p 39).

Biggeri and Giommi (1987) outline a jackknife procedure, and four associated estimators (similar to those proposed for the replication variances, see section 2.a.5.3.2). Leaver and Cage (1997) describe a jackknife approach for the US CPI where items are grouped into seven strata for the self-representing areas, each with 32 clusters and an eighth stratum for the non-self-representing areas with 12 clusters. Each replicate involves deleting one cluster in one stratum, recalculating the weights in this stratum to produce an estimate from the remaining clusters, and then using the new estimate together with the other strata (unchanged) to make a replicate estimate. Leaver and Cage (1997) point out that the stratified jackknife estimator assumes equal expected price (or price change) among the clusters within a stratum, so the resulting variances are on average expected to overestimate the true variances for items where this assumption does not hold. However, they find both under- and overestimation relative to random groups, though Shoemaker (2009) indeed finds that jackknife variance estimates are generally slightly higher than the stratified random groups method. Shoemaker (2009) and Klick and Shoemaker (2019) use a slightly different implementation of the jackknife, as a (conservative) upper bound for the variance estimates.

The bootstrap is in principle also available, with resampling of the clusters within the strata with replacement from the available clusters. The only evidence that this has been used is once again from the US where Shoemaker (2009) investigated both jackknife and bootstrap estimators as possibilities for dealing with an unusual variance estimate (see also section 2.a.5.3.2). Jackknife and bootstrap produced practically the same pattern of variance estimates, with the jackknife slightly higher than the bootstrap.

2.a.5.3.3.1. Strengths and weaknesses

The jackknife and bootstrap act in quite a similar way to the replicate samples once the sampling system has been set up in this way. Essentially they replace the choice of half-sample in each stratum. Therefore they are able to take account of the variance including such processes as imputation, product replacement and quality adjustment, as long as these are recalculated in the jackknife or bootstrap replicate. However, this is a considerable complication, and it is not clear whether it is done in the US implementations.

Otherwise the properties are rather similar to the stratified replicates approach, including the potential susceptibility to unusual observations. The bootstrap provides some additional flexibility in that over many replicates, it allows the distribution of the variance estimates to be constructed, which may demonstrate the susceptibility to outliers.

Andersson et al. (1987a) in a slightly different context find that the sampling distribution of the variance estimates over replicates is spread much wider than the replication equivalent; it is not clear how far this result can be generalised to single price indices.

2.a.5.3.4. Model-based approaches

Kott (1984), Valliant and Miller (1989), Valliant (1991) and Zhang (2010) all make use of a model-based framework for price index estimation. They all operate with (at least) the Laspeyres-type index and use an estimator of the form $\sum_i w_i p_i^{0,t}$, a weighted sum of price relatives where 0 in the superscript represents the base period and t the current period. Zhang starts with models which motivate the elementary aggregates (for Carli, Dutot and Jevons), and builds up the price relatives from these components. The motivating model for the Carli (perhaps the simplest if not the most realistic for modern implementations of the CPI) is

$$r_{ij}^{0,t} = \theta_{ti} + \eta_{tij} \tag{2.a.6}$$

with a common parameter θ for the ratio of prices between the base period 0 and t ($0 < t$), and price-specific error η with $var(\eta_{tij}) = \sigma_{ti}^2$ and $cov(\eta_{tij}, \eta_{tik}) = 0, j \neq k$. With this model the Carli is the best linear unbiased estimator (BLUE) of θ, but as Zhang points out, this does not mean that the model (1) is suitable for the price data to hand. He prefers a robust model-based variance estimator which does not depend on the distribution of η, but instead only on the property that $E(\eta_{tij}) = 0$. This estimator is

$$v_i = \frac{1}{n_i(1-n_i)} \sum_{j=1}^{n_j} \left(r_{ij}^{0,t} - \frac{1}{n_i} \sum_{k=1}^{n_j} r_{ij}^{0,t} \right)^2 \tag{2.a.7}$$

See Zhang (2010) for estimators corresponding with the Dutot and Jevons model assumptions. Zhang goes on to provide additional expressions which allow for the calculation of variances when chaining the elementary aggregates, and also for the chaining of higher-level indices.

Valliant (1991) uses a more complex model which allows for different parameters for different outlets h and correlation of prices within outlets. He also introduces an autoregressive error which allows for outlet-specific autocorrelation in time in the model errors for individual price relatives:

$$r_{hij}^{0,t} = \alpha_{th} + \omega_{thi} + \epsilon_{thij} \tag{2.a.8}$$

$$\epsilon_{thij} = \rho_h \epsilon_{(t-1)hij} + \xi_{thij} \tag{2.a.9}$$

with parameter α for outlet h, a random parameter ω for commodity i within outlet h, and with price-specific errors ϵ, with $E(\omega_{thi}) = E(\epsilon_{thij}) = 0$, $E(\omega_{thi}\omega_{t'h'i'}) = 0$, if $thi \neq t'h'i'$, $E(\omega_{thi}^2) = \sigma_{\omega_h}^2$, $E(\xi_{thij}\xi_{t'h'i'j'}) = 0$, if $thij \neq t'h'i'j'$, $E(\xi_{thij}^2) = \sigma_{\xi_h}^2$ and $-1 < \rho_h < 1$. The additional flexibility of this model suggests that it may be able to fit better, though there is no summary of the model fit in Valliant's paper (where the results are largely based on a simulated population which is nevertheless constructed from actual US price data). This approach is quite closely related to the stochastic approach to determining the appropriate form of an index number. This model leads to a range of possible (L-type) index estimators (Valliant (1991) gives seven possibilities), most of which do not correspond directly with the classical ways of constructing a consumer price index, though some of the estimators are similar. These estimators lead to different variance estimators, some of which have good design properties as well as model properties and can be interpreted as approximate design-based estimators.

As well as these explicitly model-based estimators, a number of authors have used models to estimate components of the sampling variance due to the different stages of the multistage design, particularly in the USA (Baskin, 1992, 1993, Baskin and Johnson, 1995), but also in Italy (D'Alò et al., 2006). Fixed effects (ANOVA) models can be used, but are susceptible to producing negative estimates of variance, so random effects models and restricted maximum likelihood (REML) or Bayesian estimation have been investigated. The US design has four stages of sampling, and the price relatives r can be related to the stages through a random effects model:

$$r_{ijkl}^{t,s} = \mu^{t,s} + \alpha_{i}^{t,s} + \beta_{ij}^{t,s} + \gamma_{ijk}^{t,s} + \epsilon_{ijkl}^{t,s} \tag{2.a.10}$$

where μ is a fixed effect, and α, β, γ and ϵ are mutually independent random effects with mean 0 and variances σ_{α}^2, σ_{β}^2, σ_{γ}^2 and σ_{ϵ}^2 (in principle the variances also depend on t, s, but the model is simplified so that it is not time dependent, giving time-averaged values). i, j, k and l label the sampling of PSUs, outlets, items and products respectively. These models seem to be best fitted by REML, though this does give different variance component estimates than ANOVA (Baskin and Johnson, 1995). The variance components are generally used in sample design to ensure that the greatest sample sizes are introduced at the most variable stages, but they can also be used to estimate the overall sampling error of the index, and the process is sketched by D'Alò et al. (2006, section 4).

2.a.5.3.4.1. Strengths and weaknesses

The key advantage of these model frameworks is that they do not require a probability sampling mechanism to justify the estimates; Zhang (2010) argues that this is the only realistic framework for most CPIs because they use purposive sampling. The method is also relatively simple — in this respect similar to the strictly design-based approach of section 3.1 in that setting up the model and deriving the appropriate estimators and variance estimators may involve some detailed algebra, but once it is set up it can be used relatively straightforwardly for a period. Some elements of the adjustment of the prices in an index, such as quality adjustment and imputation, are model-based, and can be included in the model structure if desired (at the cost of extra complexity). Hedonic modelling methods, which are also widely used for some items where it is difficult to account for quality differences, also fit naturally within the model-based framework (Johnson (1975) provides an early example, and there seems to have been little research on this problem since then).

The disadvantages are that it is necessary to check regularly that the model continues to fit the data. And there is the question of what is an appropriate model. The model variance captures the average difference of the observations from the model, so the choice of model makes a difference to the estimate of the variance. This makes it more difficult to justify a particular model and to claim objectivity for the variance estimates. But this has not been a particular hindrance in other parts of official statistics based on models, such as small area estimation. In this sense, Zhang's approach of using a model as a basis to motivate the existing approaches, and then calculating a robust variance, seems more likely to be widely acceptable. However, a model-based approach is not impossible, and indeed it would be possible to construct a completely model-based price index system (as suggested by Kott (1984)).

The model-based sampling error also responds in some way to von Hofsten's (1959) criticism that there is no such thing as a sampling error, in that it does not arise from the sampling process. Including all the elements that led von Hofsten to consider that sampling errors were too challenging may also be difficult in a modelling framework, but at least such a framework will treat all its elements consistently.

2.a.5.4. The development of variance estimates in the US CPI

The main push towards calculating variance estimates for the US CPI came during the 1950s, with Mudgett (1951) setting out the theory and providing a critique of BLS's compilation procedures. Adelman (1958) undertook a local study and calculated the variances of her price index, comparing with the BLS's index for the same area. Kruskal and Telser (1960) criticised the absence of studies into the variability of the CPI. McCarthy (1961) wrote a staff paper in support of a review of the CPI methods, an impressive paper which not only set out the case for calculating sampling variances for the CPI, but also made the best use of available data to calculate approximate error measures. The bias component was calculated from Swedish data, and estimates of the variance due to sampling commodities, and due to sampling both cities and outlets were calculated from detailed data made available by BLS for the review. This was the first, if rather approximate, attempt to calculate the errors for a national price index. McCarthy also set out a framework for a half-samples approach to sampling both cities and commodities which would enable variances to be estimated accounting for the complexities of price collection. He specified an additional use of both commodity half-samples within some paired cities to enable the components of the variance due to different parts of the sampling to be estimated. This approach was implemented in the US CPI from December 1963. Almost all of the subsequent development work has been undertaken by employees of BLS; although many of these have respected academic reputations, it is interesting that there has not been much direct interest from the academic community. This is, however, partly attributable to a lack of pressure from stakeholders, despite the many uses of detailed components of CPIs.

The new sampling structure led to the first substantive estimates of sampling variances for a national index – originally in some ASA conference proceedings (Wilkerson, 1964) and later in Wilkerson (1967). These demonstrated that the variances were relatively small, and suggested that focus on improving accuracy should be in other areas.

There therefore seems to have been a hiatus in the production of sampling errors, until Weber (1980) described the sampling procedure incorporating the half sample selections in detail and also set out the variance estimation methods. This was to be built into a variance estimation system, though the implementation of the calculations is not described. Weber explained that the theory for the application of half-samples was not completely worked out for complex statistics, but Kott (1983) provided a (model-based) superpopulation justification for the approach. Kott also made the first foray into designing a CPI as a model-based index (Kott, 1984), though this approach has not been taken up in a national CPI.

The model-based approach was developed further in a series of papers by Richard Valliant, who set out a model (summarised in section 2.a.5.3.4) for the evolution of prices, and used it as the basis for a series of model-based estimators (some also with good design-based properties), first under a single-stage design (Valliant and Miller, 1989), then a two-stage design (Valliant, 1991). Valliant (1999) also provided a more general view of the use of models in price indices, including the model-based approach, its relation to the stochastic approach to index numbers, and the estimation of variance components and how they are used in efficient sample allocation.

Despite Weber's description of the variance estimation system, it was not until Leaver (1990) that the next set of variance estimates, for 1978-1986 were produced, more than 20 years after Wilkerson's initial estimates. The whole price selection had been made probabilistic from the redesign introduced for 1978 (BLS, 2015, chapter 17), so this was a natural starting point. Leaver's estimates were conditional on the weights (that is, treating the weights as fixed constants), and so underestimated the total variance. They corroborated the analytical deduction of Valliant and Miller (1989) and Valliant (1991) that the standard error of the index would grow with distance from the base period, though the variances of changes in the index were approximately constant over a given lag. Leaver et al. (1991) extended the approach to unconditional variances (that is, accounting for the sampling variation in the weights), again using Taylor linearisation to combine the different variance and covariance estimates. The variances of the index level were essentially unaffected, and the variances of 12-month change increased by between 6% and 20% over the conditional variance, again in line with early results in Edgeworth (1888) on the relatively small impact of the variance of the weights. Leaver and Swanson (1992) extended the calculation of estimates to 1987-1991, taking account of the redesign from 1987 which had used variance information to rebalance the sample towards selecting additional outlets rather than additional items. These estimates also included a new component from the estimation of the base year expenditures in each replicate, which had not been included in the earlier estimates, and incorporated covariances between indices for higher level geographical areas. The estimates for this period are slightly higher than those for 1978-1986, and also more variable, but otherwise the pattern of the estimates tells much the same story.

Leaver and Valliant (1995) summarised the conditional/unconditional variance estimation work and proposed an estimator which used the replicates without the need for the Taylor linearisation which had more degrees of freedom and therefore produced more stable variance estimates. Baskin and Leaver (1996) investigated a jackknife variance estimator, and found that it had slightly larger empirical bias and empirical variance than the Taylor linearisation estimator. They also examined the use of the geometric mean in place of the Laspeyres estimator for housing, but found that this had little impact on the estimated variance. Leaver and Cage (1997) extended this to the whole CPI and to superlative indices, again finding little difference in variance estimates under the different estimators. From 1999 the CPI switched to using the geometric mean estimator, in response to the recommendations of the Boskin report (Boskin et al., 1996). Leaver & Cage also continued research into jackknife variance estimation, finding that it underestimated variance relative to the stratified random groups estimator in some item groups and overestimated in others.

One important effect identified in the series of papers by Leaver and co-authors is the impact that imputation has on the variance. This was investigated further by Leaver and Larson (2002), who calculated a set of variances using imputation across the whole index (and therefore not accounting for the imputation variance) compared with a set where imputation was done within replicates (and accounting for imputation variance). Imputation variance accounted for 0-10% of total variance in

1-month changes in fresh fruit and vegetables, but was somewhat higher for citrus fruits, which are particularly seasonal. Patterns for variance of longer-term changes were counterintuitive, and are difficult to interpret.

Decomposing the variance into components which can be used to improve the sampling has also been an important activity in the US. Several papers describe the periodic redesigns. Baskin undertook a programme of research on how these components could best be estimated (Baskin, 1992, 1993, Baskin and Johnson, 1995), preferring restricted maximum likelihood estimation, though it produced some differences compared with standard analysis of variance. It was also taken up for an analysis of the new housing sample in the CPI by Shoemaker (2002). In principle the models underlying variance components can be used to produce overall sampling errors (see section 2.a.5.3.4), but this seems not to have been done in US, where replicate errors have been available.

The US began experimentation with scanner data in the early 2000s. Leaver and Larson (2001) undertook a study on different index estimators from scanner prices of cereals, and calculated their jackknife variances. There was little difference in variances for different estimators, in line with previous results. They also calculated imputation variance for the scanner data cereals index, essentially the same as that of Leaver and Larson (2002). Here the rate of missingness was low and the imputation variance constituted only c.0.2% of total variance. Leaver and Larson (2003) repeat the variance components work of Baskin and co-authors (see previous paragraph) with scanner data and using a loglinear model; between (outlet-)chain and between item type variances are the principal sources of variation.

Developments in recent years have been more piecemeal. The US introduced a chained superlative price index in 2002, and Shoemaker (2003) investigated its sampling variance using the stratified random groups method. The variances of the superlative index were generally slightly higher than for the main index. Shoemaker (2009) and Klick and Shoemaker (2019) worked with jackknife variance estimators (see section 2.a.5.3.3), to investigate outlying variance estimates and differences between indices for different (special) populations respectively.

Around 2012 the BLS moved away from the selection of two replicates directly, and instead used a single sample. The price quotes were then randomly allocated to replicates so that variances could continue to be produced using the half-sample procedures which had been in use for variance estimation since 1964. The allocation of prices to replicates had to remain stable in order to obtain similar estimates (Shoemaker and Marsh, 2011).

2.a.5.5. Total quality estimation for CPIs

There has been a lot of work by economists on the conceptual basis of consumer prices, and the sorts of biases which arise because the concepts are not well matched by the data and methodological construction, see for example Wynne and Sigalla (1994), Moulton (1996). Dalén (1995) contrasts this with the more usual approach of survey statisticians who are concerned with sampling errors and other kinds of nonsampling errors.

2.a.5.5.1. Mean squared error estimators of CPI error

Biggeri and Giommi (1987) set out a structure for the classification of errors within a price index based on a breakdown of the mean squared error (mse) as:

$$mse(\hat{I}) = E\left[\hat{I} - E(\hat{I})\right]^2 + \left[E(\hat{I}) - I\right]^2 + (I - I^*)^2 + 2\left[E(\hat{I}) - I\right](I - I^*) \qquad (2.a.11)$$

where $\hat{I}$ is the estimated index, I is the defined goal ("true value") of the index and I^* is the ideal goal (these elements had already been described by McCarthy (1961, p 211)). The first term on the right hand side of (2.a.11) represents the total variance, including both sampling variance (as discussed so far in this paper) and measurement variance. The second term represents the (squared) bias derived from sampling and measurement. The third term represents the (squared) bias capturing how well the specified form of the index captures some ideal index number; this has often been operationalised as how well a given index approximates a superlative index, but this has been a longstanding area of debate with no universally agreed solution, and we therefore do not consider this element further here. The final term captures the interaction between the bias terms. Andersson et al. (1987b) use the same structure, with the first two terms collapsed together to form the mse of $\hat{I}$ with respect to the target index I.

Balk (1989) provides a more specific formula for the mse under the assumption that the samples for the expenditure survey, items and outlets are all independent. Dalén (1995) goes into more detail, setting out:

- two biases – one from estimation of the weights (but curiously not considering any component of variance from this estimation) and one the aggregate of the biases in the estimation of the price relatives; and

- two variances – related to two sampling stages, one from selection of item groups, and one from selection of prices.

2.a.5.5.2. Error components

Partial enumerations of the errors which potentially affect price indices have been provided by Edgeworth (1888), Bowley (1928), Morgenstern (1963, chapter 10), Biggeri and Giommi (1987), Dalén (1995) and BLS (2015). Economists have also considered many specific error sources, and these are mostly wrapped up here in "measurement error", rather than considered separately, because we focus on the statistical errors. A synthesis of these sources of error, summarising and extending the component classification of Biggeri & Giommi, and showing which among selected authors considered which errors, is shown in Table 2.a.6. It is clear that the range of potential error sources in a CPI has grown with time as there has been more detailed consideration of the processes for producing an index.

Sampling variances are only one component of the accuracy of a CPI, and there is a general feeling among authors (e.g.Wilkerson (1967), Biggeri and Giommi (1987), Dalén (1995)) that sampling errors are generally small relative to the other types of error in a price index. Specific studies on these other types of errors include:

- nonresponse error – Kersten (1985) calculates bounds on the bias in a price index induced by nonresponse in the expenditure survey which provides the weights.

- imputation error – quite large proportions of prices are not collected in any period because the products are unavailable (see for example BLS (2015), Table 3). Leaver and Larson (2002) give some examples of imputation rates and investigate the proportion of the variance due to imputation in the US CPI.

- formula error – this is a vexed question since there is no gold standard with which to compare a formula, but the general approach has been to try to approximate a superlative index formula as closely as possible, see for example Mudgett (1951, p 47-51) and Dorfman et al. (2006). In view of the continuing debate particularly in the UK about elementary aggregates, we will not consider a formula error further here.

- measurement error – many of the kinds of errors which are of concern to economists come under this category, including new items and products, and dealing with quality change, but it also includes measurement errors of the survey type resulting from the practical difficulties of the field operations. Andersson et al. (1987b) undertook a small study comparing list prices and outlet prices, which forms one component of measurement error. Dalén (1995) gave an evaluation of quality adjustment error for clothing in the Swedish CPI. Much recent attention has turned to web scraped and transaction/scanner data, and these present their own measurement problems, for example in automated classification of products, though may also avoid some of the measurement errors of observation. Little attention seems to have been directed to assessing these errors while the form of an index based on such information remains to be resolved.

- correlated price collector error. Andersson et al. (1987b) analysed this error in the Swedish CPI, and although the correlation is theoretically positive, they found many negative estimates, suggesting that the parameters were not well identified. The authors drew no conclusion about variance inflation from correlated price collector error, but noted that estimates were relatively low, <1.1 for most items.

- rounding error – a component of error which is rarely considered and rarely important in official statistics, but in the context of a CPI where both the index level and the percentage changes are regularly reported to only one decimal place, the relative error can be larger and more important. In this respect it is notable that the UK publishes 12-month inflation rates calculated from the rounded index values, in contradiction of best practice for rounding, for the purposes of presentational consistency. The only known evaluation of rounding error is by the BLS (Williams, 2006), where the error in the monthly change in an index rounded to one decimal place is of a similar magnitude to the sampling error. Rounding and sampling error also interact – for example Wilkerson (1967) says "in fact, a real change of only 0.1 per cent in the monthly CPI is significant but, since a change of this size in the published index can result from a much smaller actual change in the un-rounded figures, one cannot be sure that any particular 0.1 per cent change is significant." When the sampling error is near to or less than the effect of rounding, there will therefore be a danger of over-interpretation of changes in inflation, and the presentation of sampling error information must account for this.

These various components of error can give rise to biases and variances. It is generally challenging in a survey context to measure biases, often requiring some special study to get less biased or unbiased measures against which to evaluate. But in price indices there may be no agreement even on what the right target parameter is (see also the quotation in section 2.a.5.1.2); Dalén (1995) suggests that these biases should therefore be called bias risks, which "show ... the sensitivity of the index estimate to different, but not unreasonable, index definitions".

2.a.5.5.3. Producing an overall quality statement for a CPI

There are relatively few proposals for the theoretical form of the total error in the CPI, and only Dalén (1995) makes an attempt at numerical evaluation of a range of these variances and biases for the same index. But even he declines to combine all the biases together to give an overall impression of the mean squared error, because the estimates of the biases are themselves erratic, changing from year to year, and it is not clear that a combined version would be a satisfactory guide to the overall quality of the CPI.

Perhaps the ideal strategy to get a credible overall quality measure would be to introduce a system which produces estimates of the various quality components as part of the standard operation of a CPI. Then the variation in the quality measures could also be assessed, and a smoothed version of the total error indicator produced which would have more general application.

Combined classification, derived from Biggeri & Giommi (1987) with additions		Edgeworth (1888)	Bowley (1928)	Biggeri & Giommi (1987)	Dalén (1995)	BLS (2015)
Component A: sampling error and error variance $E[\hat{I} - E(\hat{I})]^2$						
A1	representative item sample error		(b) sampling error	A1		
A2	point of purchase sample error			A2		
A3	point of time sample error			A3		
A4	expenditure weight sample error			A4		
A5	product sample error					
A6	measurement error variance					
Component B: bias $[E(\hat{I}) - I]^2$						
B1	frame errors*			B1		
B2	coverage errors*	errors from unrepresented item groups	(d) omission of relevant classes	B2	2.2 non-represented consumption 2.7 (part) undercoverage	coverage error
B3	concept error for prices	errors in price relatives	(e) noncoincidence of the field of investigation with the objective	B3	2.8 conceptual ambiguities in certain item groups	estimation error (part)
B4	homogeneity error (of products between periods)		(a) inaccuracy of price relatives	B4		
B5	nonresponse bias			B5	2.7 (part) nonresponse	nonresponse error
B6	measurement error (includes substitution and quality adjustment)			B6	2.5 quality adjustment errors 2.6 errors in the recorded price	response error, estimation error (part)

B7	concept error for weights	errors in weights	(c) errors in weights	B7	2.1 high-level weight errors	
B8	processing error					processing error
B9	mismatch of prices and weights		(f) inappropriateness of the relative to the weight			estimation error (part)
B10	selection bias				2.7 (part) selection bias	
Component C: target bias $(I - I^*)^2$						
C1	calculated index is not target index			C1	2.3 elementary aggregate bias	estimation error (part)
C2	lack of characteristicity			C2		
C3	lack of consistency in aggregation			C3		
C4	low-level substitution bias				2.4 low-level substitution bias	

Table 2.a.6.: Synthesis classification of error sources summarised and extended from the framework of Biggeri and Giommi (1987), using selected works. * Could be split into B1a etc as a: representative items, b: points of purchase, c: households and d: products to match the component A breakdown.

2.a.5.6. Discussion and future research directions

2.a.5.6.1. Choosing a variance estimation approach

It seems possible to draw some general conclusions from the history of research on sampling variances for CPIs. Several researchers have considered the impact of variability in the estimation of the weights used in the production of price indices, generally derived from a household survey, although some countries now adjust these through the national accounting framework. There is general agreement that the effect of variability in the weights is relatively minor compared with the variability in the prices, as already noted by Edgeworth (1888).

There seem to be some clear leaders among the four main approaches described in section 2.a.5.3. Most national CPIs incorporate probability sampling in some stages of selection of prices and in the calculation of the weights, and the replication based approaches provide a natural way to estimate this variation and the variation due to non-probability but replicable parts of the procedures. This seems to be the current frontrunner, and is the only method in use for regular publication of CPI variances; variations of the method which seek to create the replicates through application of the jackknife or bootstrap seem possible, but are not in regular use, and there is some limited evidence that the variance of the variance estimate is larger with these procedures. The use of the jackknife to calculate variances for the UK CPI is explored in chapter 2.a.6.

Model-based procedures have some attractions, particularly in allowing other model-based procedures for non-response and quality adjustment to be included seamlessly, but it is less easy to explain their genesis to users, and less easy to explain what they actually mean. More work to produce and validate variance estimates using these methods is needed to provide the evidence from which to judge their usefulness.

Except in special situations, Taylor linearisation seems too complicated for practical use as a single method, although it clearly supports the use of other methods by simplifying the estimation of variances of complex statistics.

2.a.5.6.2. Presentation of quality measures

Valliant (1992) investigated the smoothing of sampling error estimates over time, and concluded that smoothed series were more in line with user expectations of the smoothness of standard errors, and had negligible impact on the coverage of confidence intervals in a simulation study. On this basis, it would seem reasonable to consider the smoothing of variance estimates, particularly for subseries where the sampling variability of the variance might be expected to be greatest.

In the US sampling errors are calculated each month with the index, but published a year in arrears for the whole year alongside the detailed CPI estimates publication (Shoemaker, 2003). These publications include a description of the sampling and the methods used to calculate the sampling errors, which probably help users (those with sufficient sophistication to understand and use sampling errors) to interpret how the measures fit with the index estimation. Reed and Rippy (2012) give an accessible introduction to the errors in the US CPI, intended for public consumption.

2.a.5.6.3. Further directions for research in quality measurement for CPIs

Further development of the model-based approach would be valuable as a way to evaluate the quality information that it can provide. The fitting and evaluation of models in the model-based approach to price indices is a first topic which needs attention. It is one of the helpful list of topics where further research on the use of models in price indices would be beneficial given by Valliant (1999), and all these areas still seem to be open for new research. It would also be interesting and important to adapt variance estimation to account for extra variance due to

- quality adjustment – suggested by von Hofsten (1959) and Leaver and Valliant (1995). This may be in part already covered by the half-sample type approaches.
- price imputation (Leaver and Valliant, 1995).
- hedonic price measurement.

There were some interesting ideas about the redesign of the Italian CPI suggested in D'Alò et al. (2006) including positive coordination of outlet samples for different commodities, which would then mean that they were no longer independent. It would be interesting to examine whether the approach of Dalén and Ohlsson (1995) could be extended to this kind of design. D'Alò et al. also propose balanced sampling of municipalities, and there are procedures for estimating the variance of balanced samples, so these could conceivably be incorporated, though this may be something which can be more straightforwardly accommodated by selecting replicates and using the balanced half sample approach. It would also be worth extending this idea, and investigating whether the variance in the CPI could be reduced through balanced sampling of municipalities/cities/areas rather than random sampling, measured through the replicate sample approach.

The US has rotation in price collection (Valliant (1991) implies 20% rotation in areas, outlets and items each year, though other work suggests that areas are replaced less frequently following each census). These rotations form panels from which even lower level subindices could be constructed. This type of structure lends itself to modelling using state space models incorporating the rotation structure, and this could be used in a time series type analysis of the variances in a way which was excluded by Baskin and Johnson (1995). In principle this approach could be used to estimate the long-term path of the variance of the overall CPI.

In short, there are many opportunities for further research on the quality assessment of consumer price indices.

2.a.6. Estimating components of the variance of the UK Consumer Price Index

2.a.6.1. Introduction

Consumer price indices have a claim to be among the most important statistics produced by national statistical offices, because of their closeness to something that everyone experiences in their daily lives, and because they have a direct impact on both income and outgoings through indexation. However, they are also one of the more complex statistics produced, with sampling in multiple stages (each potentially with stratification), and in some stages using non-probability designs (such as purposive selection) or non-measurable designs such as cut-off sampling. An outline assessment of these stages in the UK Consumer Price Index (CPI) is presented in O'Neill et al. (2017, section 8.11). In addition, the form of the Laspeyres-type aggregation of lower level indices is a ratio of sums, with each element potentially having a sampling error, which makes the calculation of quality measures for the composite output challenging. Several strategies have been proposed for producing sampling errors, sometimes in combination, and these have are reviewed in chapter 2.a.5.

There is a surprisingly long history of thought about the quality of estimates of price indices, considering that national measurements of consumer prices mostly did not start before the First World War. The first systematic investigation was undertaken by Edgeworth (1888), who already understood that the variance induced by sampling (and other) error in the weights is smaller than that induced by sampling (and other) error in the prices. Nevertheless, these lessons have not always been widely known to later workers, and similar results have been rederived on a number of occasions. Chapter 2.a.5 describes the history of the development of quality measures for price indices. Several authors have made efforts to document the different types of errors which can occur in price indices (e.g. Bowley (1928), Morgenstern (1963, chapter 10), Biggeri and Giommi (1987), Dalén (1995). The evidence from those studies which have produced estimates of sampling errors for consumer price indices at national level is that they are rather small, and therefore that other types of error are more important (Wilkerson, 1967). Morgenstern (1963) points out, however, that sampling errors may be more important in measures of change, where other errors largely cancel out.

The approaches which have been suggested for the calculation of sampling errors are Taylor linearisation, replication and model-based estimation. Replication covers a number of techniques for deriving replicates, and balanced half-samples (also known as stratified replicates) have been extensively used in the United States (US), the only country to regularly publish sampling errors for its CPI (Shoemaker, 2003); jackknife and bootstrap techniques are also potential strategies for generating replicates.

The Bureau of Labor Statistics (BLS) introduced replicate samples in its CPI operation in December 1963, and since then has gradually developed a fully probabilistic sampling system (BLS, 2015). This means that a measure of sampling variance can be both calculated and ascribed a probabilistic interpretation. The UK by contrast, in common with many other countries, has made some steps towards probability sampling (e.g. in white goods) in different stages of selection, but still has some stages which do not use probability sampling approaches. In these cases, some assumptions have to be made to make any reasonable progress, for example quasi-randomisation, assuming that a given set of data has been generated by a random procedure although probability sampling was not strictly used (Dalén, 1995). Several stages in the UK sampling procedure approximate cut-off samples, and studies of the use of these approaches in price indices have suggested that they are more accurate than random samples (De Haan et al., 1999, Dorfman et al., 2006).

Von Hofsten (1959) and Zhang (2010) argue that the sampling variance is difficult to define and calculate from the available information, particularly because there is a need to model a quasi-sampling process in stages where sampling is not randomised. Zhang therefore suggests that these problems be circumvented by calculating a model-based variance which gives a measure of the extent to which prices all move in the same direction at the same rates. This means that the result depends on the chosen model (though Zhang adopts a robust approach to minimise this dependence), whereas a sampling based approach does not have the same requirement.

Various authors have undertaken statistical quality assessments for price indices, often as studies of particular components of the quality. The impact of the sampling variability in the weights has been estimated by Balk and Kersten (1986) for the Netherlands, Biggeri and Giommi (1987) for Italy and Leaver et al. (1991) for the US. The impact of the sampling variability in the price sampling often considers only some levels of the sampling, such as Andersson et al. (1987b) for Sweden. Accounting for sampling in both weights and prices generally involves a tailored design, currently undertaken only in the US (e.g. Leaver et al. (1991)). Attempts to calculate estimates of all the quality elements including biases are even rarer, and to our knowledge only Dalén (1995) has produced a set of estimated errors, for Sweden. Although he estimated many of the components of a total error estimate, he did not combine them to produce an overall error estimate because the estimates of the biases were unstable, and it was not clear that an overall value would be a good indicator of the accuracy of the Swedish CPI.

In this chapter we estimate the sampling variability induced in the UK's CPI by two components of the sampling process, one accounting for the sampling variance in the weights, and one for the variance in the price collection. In the remainder of the chapter we recount the history of the development of sampling variance estimates in UK consumer price indices in section 2.a.6.2 (which can be compared with the wider history recounted in chapter 2.a.5), and give an overview of the design of the CPI in section 2.a.6.3. In section 2.a.6.4 we present the methods for estimating the variance of the CPI resulting from variation in its inputs, and give the results of applying these to the UK CPI. In section 2.a.6.5 we make an assessment of the relation of these results to the overall sampling error of the UK CPI. Section 2.a.6.6 presents some discussion of the implications for users of the CPI, and highlights some directions for further work.

2.a.6.2. The development of variance estimation for consumer price indices in the UK

A national consumer price index was first introduced in the UK in September 1914 in response to large changes in prices caused by the outbreak of the First World War (O'Neill et al., 2017, section 5.7). Although some of the foremost price index statisticians, who set out the theory for calculating error estimates for prices (e.g. Edgeworth (1888), Bowley (1928)), were active in the UK, they generally worked with computationally simpler systems (often producer (wholesale) prices). A new, methodologically much improved index, the Retail Prices Index (RPI), was introduced in 1956, and the first quality measures for consumer prices related to this index. Early in its history there was concern about whether the expenditure patterns were accurately measured, and several papers considered nonresponse and measurement errors in this component (for a summary see Ralph et al. (2020, section 3.1)).

The development of the RPI was supported by a Retail Prices Index Advisory Committee (RPIAC), which met irregularly. In the early 1970s RPIAC (actually mostly its Technical Subcommittee) undertook some work on the feasibility of constructing regional price indices, both for between region comparisons – which have become known as purchasing power parities (PPPs) – and for regional inflation. This led to some consideration of the sampling error in regional estimates, which were naturally based on fewer sample observations. Therefore some calculations were done on the effect of the weights and the prices (separately) on the sampling error of food prices covering about half of the food items in the basket at the time (Department of Employment, 1971, Appendix 1). Calculations were actually based on the PPP formulation, and were based on simplified calculations – the effect of the variance of the weights was approximated by calculating price indices using weights derived from three years of the appropriate survey, and treating the different years as independent (the complex design of the survey was not explicitly taken into account). Similarly the variances of the average price quotes were calculated without accounting for the complex sample collection design, and weighted together. Both standard error estimates refer to the level of the index, 0.025 due to the weights and 0.06 due to the prices (once again corroborating the deduction by Edgeworth (1888) that the variance of the prices is more important than the variance of the weights).

This work was extended by Fowler (1973), who investigated the effect of sampling error in the weights, and for the first time calculated some standard errors for the change in the index. He used successive years' weights estimates, adjusted for trends in the index, as the replicate values, and estimated that the standard error of the yearly change due to variability in the weights was between 0.0234 and 0.1022 (when the weights were derived from a survey with a sample size of 18,000 households). These values corresponded to different years, and their average could not be satisfactorily interpreted as an estimate of the standard error of the series. Fowler also considered the sampling variance of both base-weighted and chainlinked series, with the sampling errors of the long-run change in the latter due to variability in the weights being slightly greater.

After this early start there was a long gap to the next attempts to derive sampling variances for the RPI. ILO et al. (2004) summarises progress in the 1990s:

> "5.98 A number of experimental models have been tried out and calculations done for the United Kingdom. None of them has so far been acknowledged as an official method or estimate. Kenny (1995 and earlier reports) experimented with the Swedish approach on United Kingdom data. He found a standard error of the United Kingdom Retail Price Index as a whole of around 0.1, which was reasonably constant over several years, although the detailed composition of the variance varied quite a lot. Sitter and Balshaw (1998) used a pseudo-population approach but did not present any overall variance estimates."

In fact several projects to improve the quality of the RPI were underway in the 1990s, and the earlier ones are summarised in Haworth (1996). Peter Kenny's work applied a Taylor linearisation approach to estimate the variance due to location sampling, and also applied the cross-classified sampling approach of Dalén and Ohlsson (1995), estimating a standard error of 0.1 (apparently on the 12-month change) with this latter method. Unpublished work by Susan Purdon (SCPR), and Chris Skinner and Dave Holmes (University of Southampton) (summarised in Haworth (1996, paragraphs 49-57)) involved a sample optimisation exercise for the RPI using a Taylor linearisation approach to calculating variances, considering only the variances of the prices and only the element of sampling associated with outlet and product sampling (with some simplifying assumptions). As part of this project they estimated the standard deviation of the index under the existing sample allocation as 0.029, with essentially no change (despite a reduced location sample) after reallocation. However Sitter & Balshaw (1998), in an unpublished report to the ONS, noted that these estimates do not account for the finite population corrections, which may be important in some strata. Therefore, although it is appropriate for allocation, it does not give an entirely suitable estimate of the variability of the RPI.

Sitter & Balshsaw examined both the representativity and variability of the RPI using a simulated population of outlets and prices. They considered both the outlet/product sampling and the selection of items (the latter having been taken as fixed by Purdon, Skinner & Holmes). For variance estimation Sitter & Balshaw derived Taylor linearisation variance estimators, and also used a jackknife estimator. They found that the jackknife variance estimate had a lower empirical bias than the linearised estimator, and that it was easier to calculate. Later work by Sitter & Wu (in 2000 but also unpublished) used the Taylor linearisation plus jackknife approach to produce estimates of the sampling error due to outlet and location sampling of 0.049 to 0.095 index points over January 1997 and September 1997 to August 1998. They also calculated sampling errors for product (item) sampling assuming that this was done by simple random sampling (that is using the quasi-randomisation approach of Dalén (1995)), and found that the average standard error from this source was 0.19 index points. Taken together, these two estimates suggested confidence intervals of ±0.4 index points (which looks a bit narrow based on the component values), but this was generally considered to be too high for the full index because:

- it did not consider central prices, some of which were collected with zero sampling variance;
- sampling of items was assumed to be random;
- the outlet/location variance was derived using the jackknife estimator, which is known to be conservative (i.e. to overestimate the variance) on average (Wolter, 2007, p 156).

After this development work in the 1990s the focus moved to the formula effect and differences between the CPI and RPI (O'Neill et al., 2017, ch 10 et seq.). The CPI became the preferred measure of inflation in the UK, but the RPI is still produced because it has many legacy uses. A new version of the CPI including a measure of owner occupiers' housing costs, the CPIH, was introduced in 2013. It was originally assessed by the Office for Statistics Regulation to meet the criteria to be designated National Statistics; however, subsequent improvements to the methodology were identified and the designation was removed. The assessment report which eventually approved the redesignation of CPIH included a requirement to explore and publish estimates of quality (UK Statistics Authority, 2016), which has led to renewed interest in this topic.

2.a.6.3. Design of the UK Consumer Price Index

The Consumer Price Index has many inputs, but we concentrate here on the main construction pathway, which consists of weights derived primarily from the Living Costs and Food Survey (LCF) and processed through the national accounts balancing process, and prices collected from stores, the internet and from a range of other sources. Store prices include some central collections. For full details of the process of constructing the CPI see ONS (2019).

Products in the CPI are classified using COICOP (classification of individual consumption according to purpose, UN (2018)). COICOP is a hierarchical classification, and the UK CPI generally uses product strata at a level more detailed than the lowest level of the classification as the building blocks, with COICOP subclasses forming the lowest level of price indices in the system. Aggregation from the lowest indices proceeds up the hierarchy to the overall CPI using Laspeyres-type aggregation, with weights showing the importance of the component indices.

2.a.6.3.1. Expenditure weights survey design

The weights are principally derived from the LCF which has a clustered design. Postcodes sectors are chosen as the primary sampling units, and within each PSU 20, addresses are selected using a systematic sampling procedure. LCF has a responding sample of around 4700 households. Households receive an interview and also complete a diary over two weeks covering smaller expenditures, whereas larger expenditures are collected by recall covering a year. These sources are combined to give estimates of total annual expenditure on different categories of goods and services. More details of the LCF methodology are available in Bulman et al. (2017).

The weights for the CPI are derived at a broad level by balancing the survey estimates and a wide range of additional information within the household final consumption expenditure estimates from the national accounts (ONS, 2020c). More detailed breakdowns within these groups are based on the survey information or specific alternative sources for particular groups. The production of the RPI weights is simpler, because they are derived directly from the LCF and supplementary sources, without balancing through the national accounts framework. Operationally the low-level CPI weights are produced from the relative sizes of the low-level RPI weights in places where the household final consumption expenditure estimates do not provide the required level of detail.

Table 2.a.7.: Summary of sampling stages in the UK Consumer Price Index.

Sampling stage	Units
I	Locations
II	Outlets and expenditure categories
III	Representative items
IV	Products
V	Time

2.a.6.3.2. Price collection design

The sample design of the data collection for the CPI is complex, involving multiple stages of sampling, and some of these stages involve non-probability sampling. In addition, the sampling process varies for different types, according to how prices are collected, and further, the formulae used to calculate prices indices from the raw data also vary by product or item. The sample design for local price collection in the CPI, which is the most complicated part of the system from the point of view of calculating sampling errors, can be summarised as a design with five stages (summary in Table 2.a.7;

see also O'Neill et al. (2017, Box 8.1, p177)).

2.a.6.3.2.1. I – Locations

A stratified (by region) sample of locations (= defined areas based around retail 'hot spots'). The largest locations are included with certainty, and smaller locations are included with probability proportional to size (pps). Some locations are replaced each year based on the coverage of the required products. The use of pps means that the sample is self-weighting within strata. The number of sampled locations within each region is also proportional to the regional expenditure share so the regions are also self-weighting.

2.a.6.3.2.2. II – Outlets and expenditure categories

A listing process takes place in sampled locations only, to construct a frame of outlets. During listing some associated size variables are also recorded: whether an outlet is the sole outlet for a business (a "single") or part of a chain (a "multiple"), and the floor area for sales, divided into different expenditure categories (commodities) for department stores selling commodities in multiple expenditure categories. For each expenditure category a probability sample of outlets is taken, either by probability proportional to floor area or, in the case of bakers, greengrocers and butchers by simple random sampling. (The final outlet sample is therefore an indirect sample, because it results from selecting elements at a different level, but its nature as an indirect sample is not used in the index construction, so we do not pursue this line of thinking here.) So after this stage we have a sample of outlet × expenditure categories.

2.a.6.3.2.3. III – Representative items

Within each expenditure category, a small number of representative items is chosen; effectively a "representative item" is the item for which a range of prices will be observed at stage IV. So for example, the COICOP class fish covers frozen and fresh products; processed and non-processed products; and white and non-white fish. The representative items (fresh white fish fillets, fresh salmon fillets, canned tuna, fish fingers, frozen prawns and frozen breaded/battered white fish) are chosen purposively (i.e. they are not randomly selected) and are usually those with the greatest expenditure. Some representative items have their own expenditure weight (for example, bananas within the COICOP fruit class) – analogous to being in a completely enumerated stratum – whereas others represent all the other items (in this example all fruits except bananas and apples).

2.a.6.3.2.4. IV – Products

The product to price in a given store is decided by the price collector based on shelf space and their experience, so that it is one of the products with the largest sales and therefore represents a large proportion of expenditure on that representative item in that store. This is a purposive rather than random selection, and approximates most closely to a cut-off design. Dorfman et al. (2006) suggest that this is an effective design approach relative to a superlative index, based on a comparison of US and UK sampling approaches in the CPI.

2.a.6.3.2.5. V – Time

Most descriptions of sampling for CPIs do not assess differences in price sampling by time. In the US prices are collected throughout a period (in different centres), but in the UK prices have been collected on a fixed index day, a Tuesday in the middle of the month, with only minor variations in a few cases. In Council Regulation EC 701/2006 the European Union required price collection to take place across at least one working week for the HICP, or over a longer period for items with volatile prices – fresh food and energy prices. The UK already had frequent collections for energy, and fully implemented the regulation from 2018 by adding a second collection on the Friday before the main price collection day for fresh fruit and vegetable prices. UK collections are therefore systematic samples in time. We do not attempt to evaluate the variance due to time sampling here, but it would be worth further consideration, especially since two days are now sampled each month, making some variance estimation practicable.

O'Neill et al. (2017, section 8.11) give an overview of how the CPI sampling procedure reflects probabilistic statistical design considerations. The key points which affect how we may interpret variance measures in what follows are

- there is an element of cut-off sampling at stage I. Locations away from major shopping centres are always excluded. However, among the included locations, the sample is a probability sample. It is generally assumed that the excluded locations are a sufficiently small proportion of total expenditure that their exclusion has a negligible effect on the index.

- there is an element of cut-off sampling at stage II. Outlets that are outside the listing of the first 1500 outlets, are excluded from the sample. This also has a negligible impact – only in the largest locations is this a constraint. The selection of sample outlets from the listed outlets is a probability sample.

- the sampling at stage III is partly purposive, with large expenditure weights being one of the criteria for selection as representative items. This could also be considered to be a type of cut-off sampling, with items with lower weights not selected; however, in what follows we will treat this stage as if it is a quasirandom sample (Dalén, 1995).

- the sampling at stage IV is purposive, with the products which are most frequently sold being selected. This is therefore also a form of cut-off sampling, but we again treat it as if it is a quasirandom sample.

- stage V is a systematic sample of days within the month; we do not consider this element of variability in the current assessment.

So in summary, the first stages in sample selection, of locations and outlets, are approximately probability sampling stages, and the further stages are essentially cut-off samples, which we nevertheless treat as quasirandom samples. De Haan et al. (1999) and Dorfman et al. (2006) suggest that the cut-off approach may be an efficient one for price sampling.

2.a.6.3.3. Chainlinking

The CPI is a chainlinked index, with each year's price relatives (i.e. the lowest levels of the index) linked to the previous one by calculation of that index with both old and new weights every January. In other words a factor $\frac{\sum w_{y-1} I_{y1}}{\sum w_y I_{y1}}$ is applied to indices I_{yt} in year y month t, to ensure continuity of the indices at the reweighting (with summation over indices I at appropriate levels of aggregation). In fact the UK CPI has uses a double chainlinking, for weights in December and for prices in January. Since March 2017 the weights have been price-updated at both links in order to comply with EU regulations on the HICP (ONS, 2016a); in fact this procedure "brings the CPI mathematically in line with a single December link index".

2.a.6.4. Estimation of variance components

2.a.6.4.1. Variances due to the weights

The weights are derived mainly from estimates made from the LCF (see section 2.a.6.3.1). To make an assessment of the impact of the sampling variation in the weights on the quality of the CPI, we use the weights based directly on the LCF as a proxy – akin to the RPI weight calculations, but without the exclusions for high and low expenditure households (Ralph et al., 2020, section 3.1.1). We also further simplify the procedure in several ways, so in total the weights are not the same in the following ways:

- in calculating the 12-month change in the index (for the weight variances only) the weight for year y is also used for year $y-1$, on the assumption that there is little change in the weights from year to year, and therefore minimal impact of the change in weights on the variance

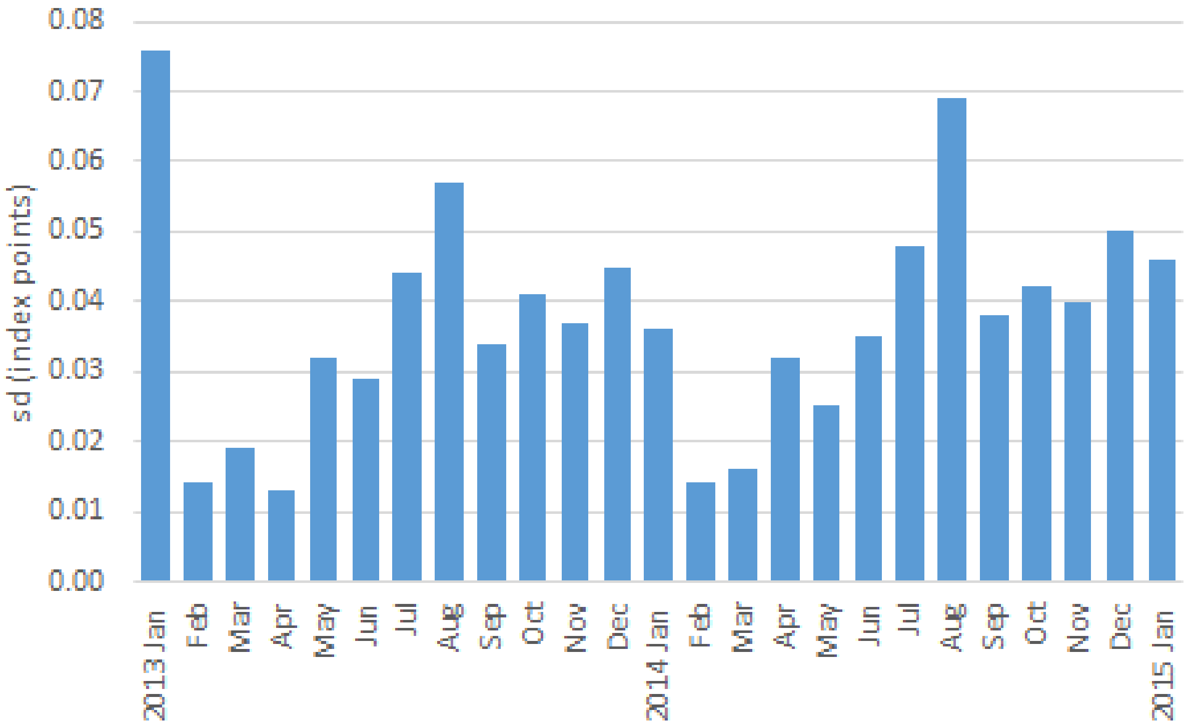
- the LCF expenditure data used in this analysis, which covers the period 2013q3 to 2014q2, spans two calendar years. Those calendar year data were used in the actual CPI for indices covering the period February 2015 to January 2016 and February 2016 to January 2017. However, in this analysis the resulting weights have been applied to indices covering Jan 2013 to Jan 2015.

- Simplifications included the use of actual LCF data without adjustments for under-reporting of alcohol and tobacco expenditure and without price-updating of the weights.

We assess the variability due to the weights using the bootstrap (Wolter, 2007, chapter 5). In each stratum of the LCF indexed by h, and containing a selected sample of n_h primary sampling units (PSUs), we select 200 samples of $n_h - 1$ PSUs from the achieved sample with replacement. In each of these 200 samples the household weights used for estimation in the LCF are multiplied by $n_h/(n_h-1)$, but not recalibrated (a small experimental reweighting of the RPI increased the estimated standard error, but by at most 0.01 index points (O'Donoghue, 2017)). The estimated RPI weights derived from these 200 replicates were used to calculate 200 replicate CPI series from January 2013 to January 2015. The variance of the CPI level was calculated as $\frac{1}{200-1}\sum_{k=1}^{200}(I_t^k - \bar{I}_t)^2$ where I_t^k is the CPI for month t calculated with the k-th replicate's weights, and $\bar{I}_t = \frac{1}{200}\sum_{k=1}^{200} I_t^k$.

The resulting estimated standard deviations are shown in Figure 2.a.11. They demonstrate the characteristic pattern of small variances at the beginning of the year when prices are little changed from

Figure 2.a.11.: Estimated standard deviations (index points) of the level of the CPI Jan 2013-Jan 2015.

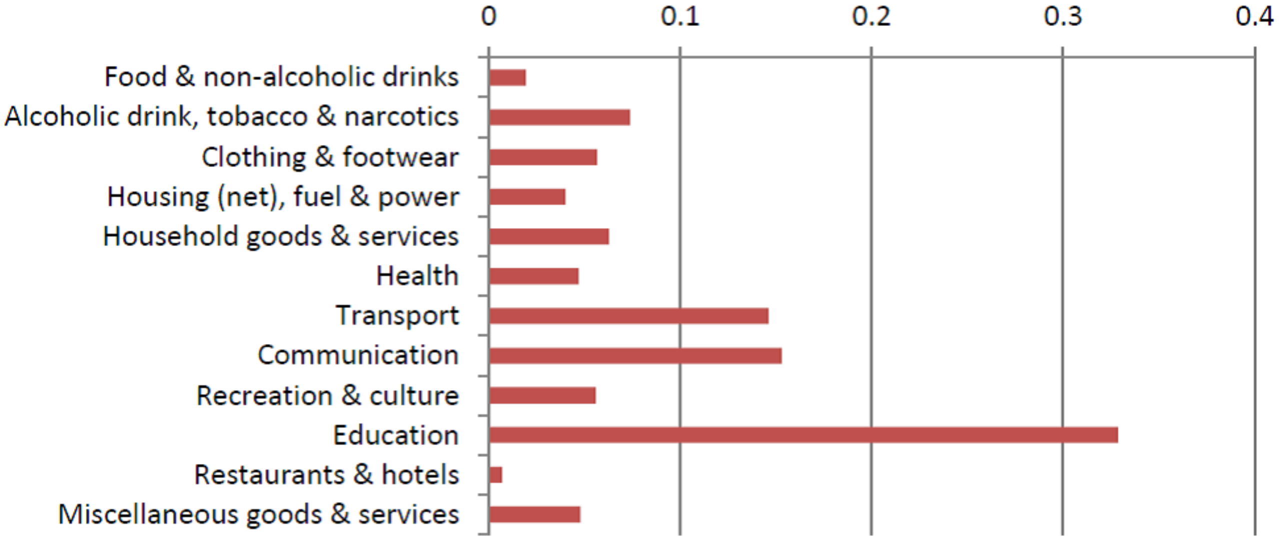


the January base prices, growing during the year, which was predicted analytically by Valliant and Miller (1989) and Valliant (1991) and has also been demonstrated in the US CPI (Leaver, 1990). The pattern is interrupted in August each year, due to the impact of air fares, which represent large expenditures made by a minority of households, and so have a large impact on the variances. Because of this pattern, the standard deviation at the end of the year approximates the standard deviation of the 12-month change in the index, so a reasonable estimate of the impact of the variability in the weights on the 12-month change in the CPI is a standard deviation of 0.05 index points.

The estimated standard deviation in January 2013 is anomalously large, and further investigation shows this to reflect the beginning of phasing in of £9000 per year university tuition fees, a large expenditure incurred by only a small proportion of households, and therefore creating a large variance. The phasing in of the tuition fees took place Oct 2012 to Oct 2014, and should therefore have had only a temporary effect on the variability of the index, so it is appropriate to discount the Jan 2013 estimated standard deviation in giving a general evaluation of the variability of the CPI. The estimated standard deviations in the COICOP division level indices are shown in Figure 2.a.12, and clearly demonstrate the variability coming from the education division.

O'Donoghue (2017) investigates the pattern of the estimated variances in more detail, giving results for different subpopulations. One interesting result is that using a more detailed breakdown of COICOP results in an increase in the estimated variance, because the additional breakdowns are less accurately

Figure 2.a.12.: Average (over Jan 2013 to Jan 2015) of estimated standard deviations due to variance in the weights for COICOP class level indices in the CPI (see section 2.a.6.3 for an explanation of COICOP).



estimated and more volatile, again through air fares but also recreation and culture (which includes computers, another large and occasional expenditure). Removing air fares and education from the index results in almost constant estimates of variances across periods. See O'Donoghue (2017) for full details.

2.a.6.4.2. Variances of prices

The sample design for the collection of prices contains multiple stages (section 2.a.6.3.2), but here we deal with stages II, III and IV, the selection of representative items and selection of products. In both cases sampling is purposive, but we treat the realised samples as if they were random (quasirandomisation).

To estimate the variances due to these two stages of the price sampling, we use the jackknife (Wolter, 2007, chapter 4), mainly for practical reasons. Around three-quarters of the CPI item indices are collected locally and processed on a central database. Of the remaining centrally collected items, around half are processed on the central database, and half in spreadsheets. The jack-knife and boot-strap calculations can both be performed relatively easily on the central database items, but determining the variance of the spreadsheet items is a time-consuming and laborious process which was most easily done using the jackknife approach. The variances from the central spreadsheets are not yet included in the analysis presented here, but will be added in the future.

2.a.6.4.2.1. Product and outlet selection

We build up the variance from the lowest level, dealing first with the selection of items. This results from selection both of outlets (stage II) and within them, products to price (stage IV) (and this cross-classified sampling is the basis of Dalén and Ohlsson (1995)'s approach to a design-based variance estimator which was followed by Peter Kenny, see section 2.a.6.2). Then we estimate the variance due to sampling representative items (stage III), combining the two on the assumption that the (quasirandom) sampling processes are independent.

For this part of the variance estimation we work directly with the 12-month change in the index. Define $p_{y,m}$ as the price quote in month m of year y, with $m = 1, \dots, 13$ to allow for the chainlinking in January (and suppressing product identifiers for readability), and $b_{y,m}$ as the corresponding base-period price. In 2014 the price relative is $r_{2014,m} = p_{2014,m}/b_{2014,m}$, but in 2015 it is $r_{2015,m} = r_{2014,13}p_{2015,m}/b_{2015,m}$ and in 2016 $r_{2016,m} = r_{2014,13}r_{2015,13}p_{2016,m}/b_{2016,m}$ in order to get a linked series of price relatives. Now dropping the time index for notational simplicity, if the 12-month price relative for product k is r_k, the number of products in stratum s is n_s, and the shop weight of the price relative is w_k^s, then the price index I_s in stratum s is the (weighted) geometric mean of the price relatives, and the equivalent index excluding product i in stratum s is

$$I_{(i)}^s = \left(\frac{\prod_{k=1}^{n_s} r_k^{w_k^s}}{r_i^{w_i^s}} \right)^{\frac{1}{\sum w_k^s - w_i^s}} . \tag{2.a.12}$$

In many cases there is no weighting information and $w_k^s \equiv 1$. The jackknife pseudovalues for the 12-month inflation rate, were then calculated as:

$$\theta_{(i)}^{s,m} = 100 n_s \left(\frac{I^{s,m}}{I^{s,m-12}} - 1 \right) - 100(n_s - 1) \left(\frac{I_{(i)}^{s,m}}{I_{(i)}^{s,m-12}} - 1 \right) \tag{2.a.13}$$

and the variance for stratum s was calculated using the standard estimator (Wolter, 2007, equation (4.2.3)):

$$\hat{v}_1^s = \frac{1}{n_s(n_s-1)} \sum_{i=1}^{n_s} \left(\theta_{(i)}^{s,m} - \bar{\theta}^{s,m} \right)^2 \text{ where } \bar{\theta}^{s,m} = \frac{1}{n_s} \sum_{i=1}^{n_s} \theta_{(i)}^{s,m} \tag{2.a.14}$$

which gives the sampling variance for a with-replacement sampling scheme, without further adjustments to account for the finite population (Wolter, 2007, section 4.3-4.4). In most cases of products within outlets this is entirely reasonable as the number of outlet-product combinations is potentially large. For centrally compiled prices quotes where the number of combinations is sometimes much smaller, if the outlets and products priced were judged to represent the complete population, the variance was set to zero. For some items, this assumption does not strictly hold, for example gas and electricity tariffs, where prices are not collected from some of the smallest suppliers. This means that the total variance is underestimated, but this element is likely to be small and is judged to be unlikely to have a material effect on the estimate of the overall variance.

Once the stratum level variances had been calculated using equation 2.a.14, they were aggregated to give variances (for stages II and IV) for items, COICOP sub-class, and higher level aggregates. The process was repeated for each month from January 2015 to January 2017 (where the period is the end of the 12-month change – the equivalent of its publication month).

2.a.6.4.2.2. Representative item selection

Then we turn our attention to stage III, again treating the observed sample of representative items as a quasirandom sample. We distinguish two situations:

1. where the selected representative items have their own weight (for example bananas are weighted according to proportion of expenditure on bananas) or, equivalently, where the representative items cover the whole of a category (for example air fares). These effectively act as if they were in a completely enumerated (CE) stratum at stage III and do not contribute to the variance. They are assigned a representativity indicator $q_j = 0$, where j indexes the item.

2. where the selected representative items also take the weight of other (not selected) items, and therefore act as sampled (not CE) items, in which case we set $q_j = 1$.

At this level of the sampling we need the inflation rate from the Laspeyres-type aggregation of the indices corresponding to the representative items. Denote the index for a representative item by J_j for product j, which has already been aggregated from the individual price relatives and chainlinked. Then within each COICOP category c we have the Laspeyres-type index L, and jackknife using only indices for which $q_j = 1$ (which has the effect of attributing no variance to the CE parts, which are the same in every jackknife replicate). The drop-one index is

$$L^c_{(j)} = \frac{1}{\sum_{i=1,i\neq j}^{n} w_i} \sum_{i=1,i\neq j}^{n} w_i J_i \text{ for } \{j : q_j = 1\}. \tag{2.a.15}$$

Then as before we calculate a jackknife pseudovalue

$$\tau^{c,m}_{(j)} = 100 n_c \left(\frac{L^{c,m}}{L^{c,m-12}} - 1 \right) - 100(n_c - 1) \left(\frac{L^{c,m}_{(j)}}{L^{c,m-12}_{(j)}} - 1 \right) \tag{2.a.16}$$

and use the same jackknife variance estimator (2.a.14) with these inputs

$$\hat{v}^c_1(c) = \frac{1}{n^*_c(n^*_c - 1)} \sum_{j=1}^{n_c} q_j \left(\tau^c_{(j)} - \bar{\tau}^c \right)^2 \text{ where } n^*_c = \sum_{j=1}^{n_c} q_j \text{ and } \bar{\tau}^c = \frac{1}{n^*_c} \sum_{j=1}^{n} q_j \tau^c_{(j)}. \tag{2.a.17}$$

It was not possible to calculate variances in some cases at each sampling phase, either because

(a) there were fewer than two locally collected price relatives in a stratum (particularly where stratification was at the most detailed region × shop type level)

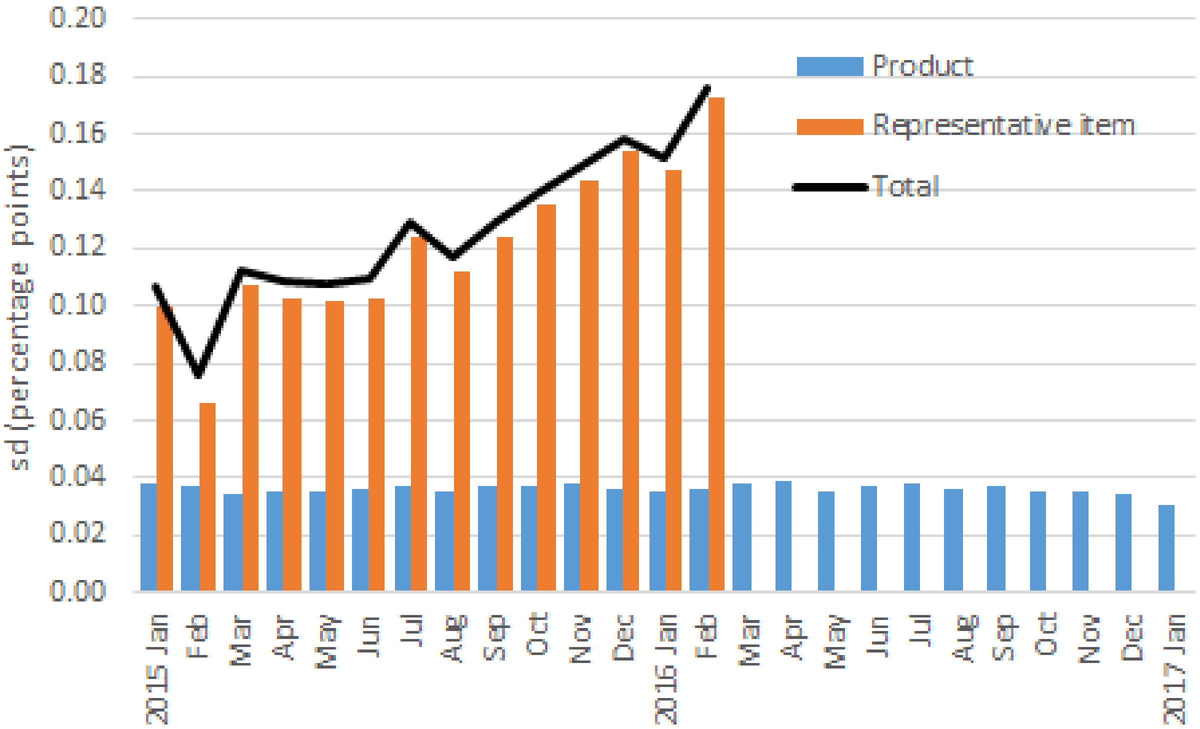
(b) the representative items had only been included in the basket since the start of the year

In these cases, the stratum variance was set equal to the weighted mean of the other non-zero stratum variances for the same item or the same 5-digit COICOP category (as appropriate) in the same month.

2.a.6.4.3. Combining representative product and outlet-item selection

The total variance for stages II-IV of sampling of prices in the CPI, $\hat{v}$, was then estimated by summing the component estimates $\hat{v} = \sum_c \hat{v}_1^c + \sum_s \hat{v}_1^s$, since these sample selections are independent. The results, with the representative item and product sampling variance components, are shown in Figure 2.a.13.

Figure 2.a.13.: Estimated standard deviation (measured in percentage points) of the 12-month per cent change in CPIH. The period plotted is the end of the 12-months – the equivalent of the publication month for the 12-month change.

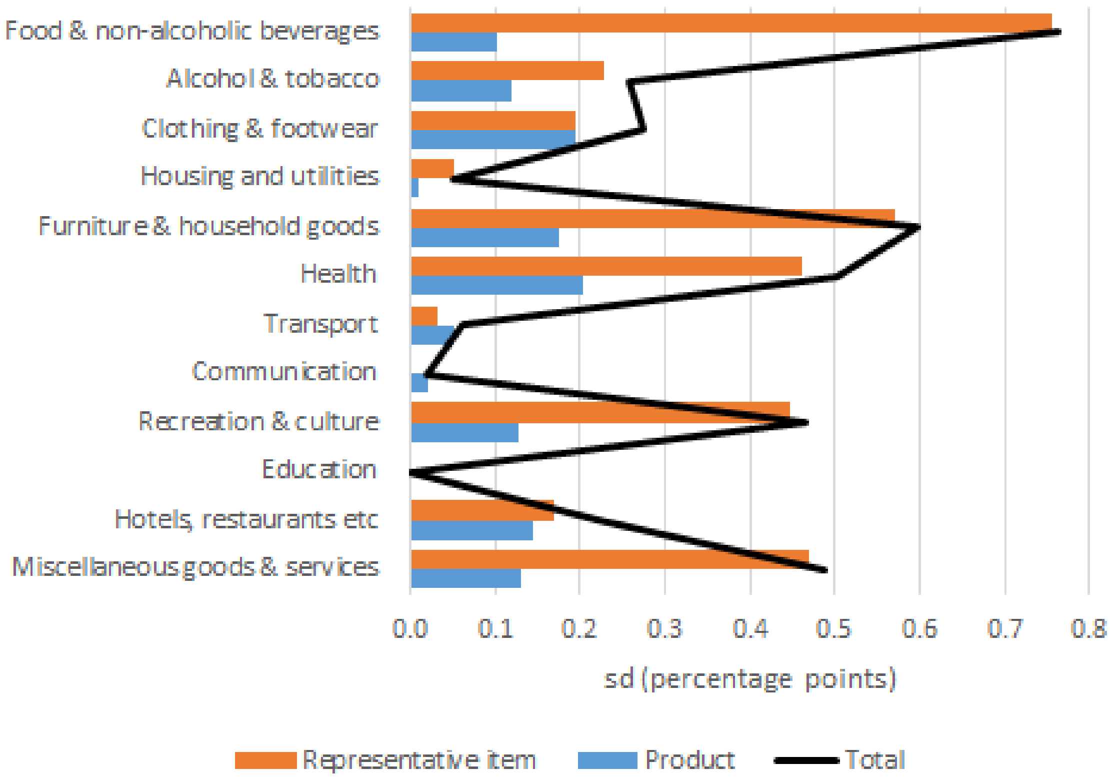


The total standard deviation of the 12-month rate of change for CPIH due to stages II-IV of the price sampling ranges between 0.076 and 0.176 percentage points over the period examined, and again shows the characteristic pattern of growth with distance from the base period in the representative item sampling, as noted by Valliant and Miller (1989) and Valliant (1991), though the variance from the product-outlet sampling is relatively constant. Most of the variance comes from the representative item sampling.

There are considerable differences between the estimated variances within the COICOP divisions (Figure 2.a.14). Division 10, education, has the highest overall variance, derived entirely from product variation, and is mainly due to the variation in costs of part time education classes at local colleges (though this is not yet shown in figure 2.a.14 pending some revised calculations). This is a category which has few items, and can be affected by one item which happens to be volatile. The biggest variance from representative item selection comes from COICOP divisions 1 (food and non-alcoholic beverages) and 5 (furniture and households goods), and reflects their diverse range of products, including (in division 5) glassware, tableware, small and large electrical tools and appliances for home and garden.

At the other end of the scale, housing and utility costs, COICOP division 4 has an overall standard deviation of 0.05 per cent. This is a division that contains many central items, and many of these items represent themselves, and therefore have little or no product variance. The price indices for rental equivalents and private rentals fall within this division. Between them, they have a weight of over 20 per cent of the total CPIH basket but they have been assigned zero variance, because they are produced by the Valuation Office Agency and their accuracy has not been estimated. The underlying data are however available, so this component could be estimated, but this is left for further work. Their absence is unlikely to have too great an effect on the overall estimate of the precision of CPIH because of the large number of rentals that are used in the calculations and because these are indices that tend to evolve fairly slowly.

Figure 2.a.14.: Average (over Jan 2013 to Jan 2015) of estimated standard deviations (measured in percentage points) of the 12-month per cent change in CPIH for COICOP classes.



2.a.6.5. A view of sampling error in the UK CPI

2.a.6.5.1. Error in the CPI

The investigations reported here provide useful insight on the variability of the CPI, but have used a series of approximations and assumptions. Some of these are likely to cause the variances to be overestimates of the actual variability in the index, and our subjective ranking of these in order of importance is

O1 use of LCF-based weights instead of weights based on HFCE. The HFCE weights, which have been adjusted for known underreporting in some commodities, and processed through balancing, are likely to be considerably less variable than the weights based on a single LCF survey year.

O2 when calculating the within item variance it is assumed that both the outlets and the specific products are chosen randomly. This is likely to lead to an over-estimate of the within item variance, as central and regional central shops are selected with certainty, and price collectors are asked to select the most sold item, all leading to less variability in the actual sampling procedure than is assumed (but see also U2 below).

O3 jackknife variance estimates on average slightly overestimate the true variances (Wolter, 2007, p 156-8).

There are equally simplifications and assumptions which mean that the variance estimates presented here are likely to underestimate the total variance in the CPI. We again present a subjective ranking:

U1 The variance component for location selection (stage I in the sampling) is missing.

U2 The sampling of the most-bought products means that rarer products, amongst which are the more extreme prices, are not included in the sample, reducing the variance relative to a random sample of the whole population.

U3 Weights within COICOP subcategories (i.e. item weights and stratum weights) were held fixed.

U4 Some centrally collected prices are assigned zero variance when a small part of the population remains unsampled (as in the gas and electricity example in section 2.a.6.4.2).

U5 The variance component for owner occupiers housing costs and rental costs is missing.

Some further approximations may work in either direction; the main one is the imputation of variances where they could not be calculated.

We have no way to assess the relative impact of these different components of under-, over- and mis-estimation of the variances overall, but note that each in isolation can be reasonably be judged to have a small impact on the variance estimates which we have already been able to make. So, given that there is a certain amount of offsetting in what are already small differences in the variances, it does not seem unreasonable to assume that the estimated variances provide a realistic assessment of the accuracy of the CPIH. That is that the overall standard error in the 12-month inflation rate due to sampling prices is around 0.12 percentage points, and the corresponding standard error due to the weights is around 0.05 percentage points. If we conservatively take these to be independent, this leads to approximate confidence intervals for this change of ±0.26. We note that the price variance is not much different to the value obtained by Sitter & Wu (see section 2.a.6.2), but this is not so surprising since the methodology adopted was essentially the same.

Dalén (1995) attempts to estimate the overall quality of the Swedish CPI, covering a range of error components. Some of the bias risks that he considers, for example estimates from different approaches to measuring owner occupiers' housing costs, can already be assessed in the UK from previous research work (ONS, 2020d). However, bringing this information together and combining it with the estimates of sampling errors is not straightforward, and is left for future research.

2.a.6.6. Discussion

The estimation of the sampling error of a price index is a complex task, because the inputs arise from surveys with complex designs, and the price collection in particular contains many stages, not all of which involve probability sampling (at least in the UK CPI). We have nevertheless taken a design-based view of the sampling error of the UK CPI, and attempted to estimate the variances due to the different inputs separately. The US approach is more holistic, with the whole index calculation system designed to deal with replicate samples, so that all the stages of sampling variation can be dealt with together (see chapter 2.a.5 for a review), and an alternative approach would be to create pseudoreplicates from the UK system and to use these to calculate the sampling error in one step. This would have the advantage of allowing the incorporation of some of the elements of sampling which have been omitted in our analysis here, particularly the sampling of locations.

Further exploration of the model-based variance approach suggested by Zhang (2010) would also be worthwhile. It would provide an alternative view of the variance of the CPI (and the comparison with the design-based variances might also generate some interesting insights), and has the further advantage of potentially allowing the incorporation of some uncertainty from model-based procedures used in consumer price index construction, such as imputation and hedonic modelling.

2.b. Estimation of local spatial price indices using scanner data

2.b.1. Introduction

Over the last decade, there has been a growing interest in using scanner data on retail prices for constructing both temporal consumer price indexes and sub national Spatial Consumer Price Indexes (SN-SCPIs). This report refers to the second field of research.

The availability of high-frequency "scanner data" in addition to other sources of data enables price statisticians to deal with the SN-SCPIs issue from a renewed approach. These data benefit from an impressive coverage of transactions along with information on: sales; expenditure; quantities; and quality with very detailed information on characteristics of products sold (brand, size and type of outlet) provided at barcode level or, more precisely, the GTIN (Global Trade Item Number) code. The scanner data of the modern distribution can provide millions of prices for thousands of products (GTIN code). They predominantly refer to supermarkets and hypermarkets, especially for food, beverages and personal and home care products. After a process of data cleaning and trimming outliers, unit value price per item code can be computed by dividing total turnover for that item by the total quantity sold.

Regardless of which provider scanner data come from, NSIs (National Statistical Institutes) must reclassify them in order to make them suitable for constructing the mentioned indexes. It should be noted that the SN-SCPIs are in essence direct spatial price level comparisons, because within a country, there is a common currency.

The unit of research implemented two experiments by using a scanner data base provided by Istat (Italian National Statistical Institute), thanks to an agreement between Istat and Dagum Centre.

As it will be better explained in the following sections and , the two implemented experiments use the same data base and elementary data, but differ in the application of the principles of comparability and representativity and in the methods of construction of the SN-SCPIs.

In the first experiment, the principles and the construction procedure is quite similar to the one used in the ICP (International Comparison Program) of the World Bank to compute the international PPPs (Purchasing Power Parities). In particular, the principle of comparability is applied in a very tight way by considering the comparisons of the "like to like" items (products) for the different sub-national areas. In this case there is the risk that not all the products are available in all the areas. However, the information on the turnover of each product allows to give a weight to each product within every area. Under this approach the lowest level of aggregation of the products is the the Basic Heading (BH) level, as done by the World Bank.

The second experiment follows a complete different approach. The principle of comparability is applied at the level of group of products, by loosing the specifications of the elementary products. The approach considers the unit value prices from the consumer side (or point of view). The hypothesis is that the elementary products (items) belonging to each group satisfy in any case the same consumer needs (and may be gives him the same utility), also if the brands, quality, etc..are different. The comparison is therefore done by considering the average level of prices of the group of products purchased in the different areas, considering the basket of elementary products that the consumers of each area have really purchased. Then the average level of prices of the group of products is aggregated to obtain the SB-SPIs for each sub national area. Therefore, these groups, and not the BHs, are the building blocks of the comparison, defined using the ECOICOP-8-digit classes of products.

Both experiments estimate the SPIs at provincial level (NUTS 3 level in EU classification). Moreover, in order to calculate SPIs closer to the prices paid by the poor, preliminary experiments have been conducted by using the data of the first quintile of the price distributions, assuming that the poor purchase the cheaper items of a product.

The specification of the work done follows. In section 2.b.2, information on the characteristics of the available scanner data base is presented, highlighting the advantages and the disadvantages in using them for the computation of the SN-SCPIs. In sections and , the procedure, methodology, and the main results of the estimation of the indexes, by using the above mentioned approaches, and also with reference to prices of the cheaper items, are presented. As the results obtained by the two approaches at aggregate level are quite different, in section 2.b.5, we propose some explanation of the differences, also if the two approaches are based on different hypotheses, and difficult to compare at aggregate level. Section 2.b.6 is dedicated to the exploration of the places where people in condition of absolute poverty purchase some large consumption products. Finally section 2.b.7 describes how to measure the impact of cost of living of the poor on the estimation of local poverty rates. Final remarks and recommendations in section 2.b.8 conclude the report.

2.b.2. The scanner data base: advantages and limitations for the computation of sub-national spatial consumer price indexes

Since 2014, Istat (the Italian National Statistical Institute) received the scanner data on retail prices by the market research company ACNielsen that is authorised to do it by the chains of modern distribution in the framework of an agreement with the Association of Modern Distribution. ACNielsen provides Istat with scanner data on a weekly basis by uploading the data files on a dedicated Istat web portal.

Istat was interested in using these scanner data on retail prices for constructing both temporal consumer price indexes and sub national Spatial Consumer Price Indexes (SN-SCPIs). Scanner data has been recently introduced by Istat in the official CPI computation, while until now they have been used in the construction of the SN-SCPI only in an experimental way.

These data benefit from an impressive coverage of transactions along with information on: sales; expenditure; quantities; and quality with very detailed information on characteristics of products sold (brand, size and type of outlet) provided at barcode level or, more precisely, at the GTIN (Global Trade Item Number) code. The scanner data of the modern distribution provide millions of prices for thousands of products identified by the GTIN code. However, as already underlined in the introduction, only after the process of data cleaning and trimming outliers, unit value price per item code can be computed by dividing total turnover for that item by the total quantity sold.

More recently, Istat has initiated a series of experiments on the estimations of SN-SCPIs at a regional level on an annual basis[1]. In October 2018, an agreement between Istat and ASESD-Dagum Centre was signed to implement the tasks of the Makswell project.

To do these activities, Istat provided to the ASESD-Dagum Centre the scanner data base referred at the years 2017 and 2018. The data base refers to a random sample of approximately 1,800 outlets, hypermarkets (more than 500) and supermarkets (almost 1,300), and contains data concerning the grocery products sold in the most important retail chains (95% of modern retail chain distribution that covers 55.4% of total retail trade distribution for this category of products). More specifically, scanner data for 1,781 outlets of the main 16 Retail Trade Chains (RTCs) covering the process to entire national territory. Outlets have been stratified according to provinces (107), chains distribution (16) and outlet-types (hypermarket, supermarket) for a total of more than 800 strata. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. The research group decided to use the 2018 data base for the analysis. For each GTIN, prices were calculated taking into account turnover and quantities: weekly unit value price is equal to the weekly turnover divided by weekly quantities. Monthly and annual unit value prices are calculated by the arithmetic mean of weekly prices weighted with quantities.

Taking account of the results of the experiments and of the many discussions among the members of ASESD-Dagum Centre and the researchers staff of the Istat' Price Statistics Unit, we can summarize here the various advantages and some limitations in using the scanner data for the computation of the SN-SCPIs.

The main advantage is that scanner data may help to overcome the issue of price data availability in the various areas involved in the comparisons by fulfilling the requirements of representativeness and comparability that emerge when compiling SCPIs. Due to the high territorial coverage which characterizes scanner data, we are able to compare price levels among the various geographical areas within a country. In addition, it is worth noting that GTIN codes describe the products in detail and they are generally the same for each item at national level. In this way, we can solve the issue of comparability. Since detailed information on turnover and quantities for each item code in every area is available, it is possible to account for the economic importance of each item in its own market, thus fulfilling the representativeness requirement. Moreover, as different modern RTCs can sell products of different quality and offer additional services, information on the type of outlet and retail chain can be included in order to account for these quality characteristics that may influence the price of a product. Moreover, if some products acquired by consumers with reduced quantities, the availability of turnover weights (defined considering sampling weights) allows us to correctly include the corresponding representativeness of these products in terms of the total turnover of the BH or the group at which the products in question belong.

[1] The experiments were conducted in subsequent improved versions of the scanner data base constructed for experimental CPI computation, analyzing the potential advantages and the empirical issues deriving from the use of the scanner data to estimate the SPIs (Laureti and Polidoro, 2017, Laureti et al., 2017).

Other advantages of the use of scanner data are: (i) the reduction of measurement errors. By using as price concept the unit value for each GTIN we can refer to a more accurate measure of an average transaction price than an isolated price quotation Diewert (1995) as in the case of traditional price data collection; (ii) The reduction of conceptual uncertainty. The GTIN unit value is a more representative price paid by consumers over the reference period than the usual price collected using traditional on-field surveys. These prices include temporary price promotion and reflect the actual price paid by consumers. Moreover, by aggregating over a year it is possible to smooth out the effect of price and quantity bouncing behavior; (iii) using scanner data we add time dimension to multilateral spatial price comparisons since detailed data are usually available at the point of sale and at the time of transaction. Another advantage (iv) is the use of itemized information contained in scanner data. Indeed, when using the unit value approach, items must be tightly defined at a fine level of aggregation to maximise homogeneity and prevent quality differences from affecting the unit values. Finally, (v) it is obvious that using scanner data to carry out spatial comparisons will increase cost efficiency since price data collection may be limited to traditional stores and shops thus lowering data collection costs for the NSI.

However, some limitations must be noted in the context of this study. The available scanner data: (i) do not cover all the retail chains of modern distribution; (ii) practically cover almost all the provinces (103 over 107), but the rural areas are not covered; (iii) cannot be used for perishables and seasonal products such as vegetables, fruit and meat, and fresh fish, since these products are sold at price per quantity and are not pre-packaged with GTIN codes.

Moreover, we have to consider that, in any case, all the scanner data available cover about the 10,5% in terms of the total expenditures of families for the consumption. In addition, this share is not uniform across the Italian territory.

Therefore, it is evident that to estimate a complete set of SN-SCPIs, it is necessary to build up a data base that could allow the estimation of these indexes related to the entire universe of household consumption. In fact, Istat collects consumer prices by using different sources: territorial surveys at the outlets by non-probability samples; use of administrative data, use of scanner data (Big data). Therefore, a strategy to use and integrate all the consumer price sources of data must be followed, considering also the fact that we need to face the issue due to the fact that the data come both from probability and non-probability sample. Istat is working on this line.

For the time being, this unit of research implemented two experiments to compute SN-SCPIs at provincial level in order to use them to adjust the local economic poverty indicators taking into account of the differences in the cost of living in the different areas by using a scanner data base. The first experiment is conducted following in tight way the principles and procedure followed by the ICP (International Comparison Program), coordinated by the World Bank, to compute the International PPPs (purchasing Power Parities). The second experiment has been conducted, by using a different innovative approach regarding the definition of the principle of comparability to verify its validity. Moreover, in order to calculate SPIs closer to the prices paid by the poor, preliminary experiments have been conducted by using the data for the first quintile of the price distribution, assuming that the poor purchase the cheaper items of a product.

The procedures followed and the results obtained are presented in the following sections.

2.b.3. Spatial price indexes: World Bank approach

In this experiment we followed the principles and the construction procedure similar to that used in the ICP (International Comparison Program) of the World Bank to compute the international PPPs (Purchasing Power Parities) at the level of the BHs. For this reasons we provisionally consider the computed SPI as PPPs.

Indeed, a two-step procedure is adopted (World Bank (2015)). In the first step, provincial PPPs are computed at BH level using the Country Product Dummy (CPD) model (which is denoted as a group of similar well-defined goods or services) by comparing price and quantity data referring to products sold in the various Italian provinces while in the second step, we aggregate the results from BH level comparisons to higher level aggregates (food and non-food products) using the GEKS procedure based on Fisher indexes.

In this report, we use the 2018 dataset that does not include a cut-off line for the products. Therefore, by considering all the products, identified by the corresponding GTIN, we can include in price comparisons not also the best-seller products but also those products acquired by consumers with reduced quantities. The availability of turnover weights (defined considering sampling weights) allows us to correctly include the corresponding representativeness of these products in terms of the total turnover of the BH at which the products in question belong.

2.b.3.1. Aggregation method at BH level: BH PPPs

As underlined above, scanner data bring detailed information about the characteristics of the elementary product and information about turnover of that specific product, allowing the comparison of "like to like" products. Weights for each specific product are based on turnover for that product.

For each GTIN, weight is obtained dividing weighted turnover for the total by weighted turnover for that product for each province.

Since product overlaps exhibit a chain structure, the weighted CPD method exhibits some aspects of spatial chaining and therefore we selected this method for computed provincial PPPs for product aggregates. With the aim of taking into account the economic importance (representativeness) of each product expressed by expenditure weights w_{ijr} based on turnover, we used a weighted CPD model. In this way, the representativeness requirement can be achieved by computed weighted spatial index numbers.

Let us assume that we are attempting to make a spatial comparison of prices between Mr provinces, with $r = 1, 2, \ldots, R$ Regions. In the first stage of aggregation of price data at item level, which leads to price comparisons at BH level, p_{ij} and q_{ij} represent price and quantity of $i - th$ item in $j - th$ province $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, Mr$. In order to compute provincial PPPs, we used as already mentioned the CPD model according to the approach followed by the World Bank. Besides accounting for quality variations in the cross-area price data, CPD is a regression-based econometric methodology that can be extended and generalized in order to provide a comprehensive framework for carrying out both international and intra-national. The literature is still expanding and a recent paper by Rao and Hajargasht (2016) further developed the CPD-based stochastic approach through the use of modern econometric tools. This method suggests that price levels are estimated by regressing logarithms of prices on provinces for each province and product dummy variables; the model is given for each BH by:

$$\begin{aligned} lnp_{ij} &= lnPPP_{j} + lnPPP_{i} + ln\mu_{ijr} \\ &= \pi_j + \gamma_k + \upsilon_{ij} \\ \textstyle\sum_{j}^{M_r} \pi_j D^j &+ \textstyle\sum_{i=1}^{n} \gamma_i D^i + \upsilon_{ijr} \end{aligned} \tag{2.b.1}$$

where D^j is a provincial-dummy variable that takes value equal to 1 if the price observation is from $j-th$ province; and D^i is a $i-dummy$ variable that takes value equal to 1 if the price observation is for $i-th$ commodity. The random disturbance is assumed to satisfy the standard assumptions of a multiple regression model. In order to estimate parameters of this model, we impose normalization $\sum_{j}^{M_r} \pi_j = 0$ thus treating all provinces in a symmetric manner. If $\widehat{\pi_j} = (1, 2, ...M_r)$ are estimated parameters, PPP for the province j in region r is given by $WR\_PPP_j = e^{\widehat{\pi}_j}$. The CPD method based price comparisons are transitive and base-invariant. With the aim of taking into account the economic importance (representativity) of each product expressed by expenditure weights w_{ijr} based on turnover we used a weighted CPD model:

$$\sqrt{w_{ijr}} lnp_{ijr} = \sum_{j=1}^{M_r} \pi_j \sqrt{w_{ijr}} D^j + \sum_{i=1}^{n} \eta_i \sqrt{w_{ijr}} D^i + \sqrt{w_{ijr}} \tag{2.b.2}$$

2.b.3.2. Aggregation above BHs: Provincial PPPs

The next and final step for compiling provincial price comparisons is to aggregate the results from BH level comparisons to higher level aggregates. Let us assume that there are L basic headings $(l = 1, \ldots, L)$ and e_i^r expenditure for $i-th$ BH in province r. We decided to use the Fisher price index since it has a range of axiomatic and economic theoretic properties. The Fisher index is given by:

$$P_{rk}^{Fisher} = \sqrt{P_{rk}^{Laspeyres} \cdot P_{rk}^{Paasche}} \tag{2.b.3}$$

Where:

$$P_{rk}^{Laspeyres} = \frac{\sum_{l=1}^{L} p_l^k q_l^r}{\sum_{l=1}^{L} p_l^r q_l^r} = \sum s_i^r \left(\frac{p_l^k}{p_l^r}\right) \tag{2.b.4}$$

$$P_{rk}^{Paasche} = \frac{\sum_{l=1}^{N} p_l^k q_l^k}{\sum_{l=1}^{N} p_l^r q_l^k} = \left[\sum_l s_l^k \left(\frac{p_l^k}{p_l^r} \right)^{-1} \right]^{-1} \tag{2.b.5}$$

with:

$$s_i^r = \frac{e_i^r}{\sum_{l=1}^{L} e_l^r} = \frac{p_l^r q_l^r}{\sum_{l=1}^{L} p_i^r q_l^r} \tag{2.b.6}$$

As the Fisher binary index in eq. 2.b.3 is not transitive, it is possible to use the procedure suggested by Gini (1931),Elteto and Koves (1964) and Szulc (1964b) referred to as the Gini-Elteto-Koves-Schultz (GEKS) index to generate transitive multilateral price comparisons across different regions. The resulting index is given by:

$$P_{rk}^{GEKS-FISHER} = \prod_{r=1}^{R} \left[P_{rs}^{Fisher} \cdot P_{sk}^{Fisher} \right]^{1/R} \tag{2.b.7}$$

The GEKS-Fisher based formula is used in cross-country comparisons made within the ICP at the World Bank (2015) and the OECD-Eurostat comparisons. In order to obtain a set of R_PPP_s that refers to the group of regions (Italy) we standardized the GEKS-Fisher based PPPs (S-GEKS).

As these PPPs are now transitive, the ratios between the PPPs for each base are the same. In order to obtain a set of PPPs that has the group of countries as a base – thereby ensuring a neutral presentation - it is necessary to standardise the PPPs in the matrix. This is done by dividing each PPP by the geometric mean of the PPPs in its column.

2.b.3.3. Results

In order to compile a spatial price index for the 103 Italian provinces at BH level, the dataset used includes 2,032,574 annual price quotes concerning 72 BHs for a total of 63,256 products (GTIN code).

To improve the quality of price comparisons, defined by the strength of interconnections and overlaps in the priced items across different provinces, the following group of products: whole milk, low-fat milk, olive oil, aged cheese, other cheese, lager beer and frozen seafood were excluded, since in these cases price data does not exhibit spatial chain. Table 2.b.1 reports the 65 BHs included in the analysis which corresponds to the ECOICOP 5 digit with a number of different products included in each BH.

Table 2.b.1.: BHs included in the analysis

Sub-classes	Description	number of products
01.1.1.1.0	Rice	574
01.1.1.2.0	Flour and other cereals	1056
01.1.1.3.2	Bread	729
01.1.1.4.2	Pastry products	7023
01.1.1.4.3	Bakery products	3227
01.1.1.6.1	Dry pasta	3833
01.1.1.6.1	Fresh pasta	677
01.1.1.6.2	Pasta preparations	1446
01.1.1.7.0	Cereal for breakfast	652
01.1.2.7.2	Dried, salted or smoked meat	1781
01.1.2.8.2	Other meat preparation	546
01.1.3.2.0	Frozen fish	307
01.1.3.6.0	Other fish and fruits	1534
01.1.4.3.0	Preserved milk	692
01.1.4.4.0	Yogurt	2122
01.1.4.5.2	Cheese and curd	1660
01.1.4.6.0	Other milk products	1120
01.1.4.7.0	Eggs	524
01.1.5.1.0	Butter	343
01.1.5.2.0	Margarine and other vegetable fats	32
01.1.5.4.0	Other edible oils	332
01.1.6.3.0	Dried fruits and nuts	2183
01.1.6.4.0	Preserved fruits and fruit-based products	616
01.1.7.2.0	Frozen vegetables	498
01.1.7.3.2	Vegetables in packs	1450
01.1.7.3.3	Dried vegetables	606
01.1.7.3.4	Potatoes	2353
01.1.7.3.5	Vegetable-based preparations	90
01.1.7.4.0	Frozen potatoes	92
01.1.8.1.0	Sugar	163
01.1.8.2.0	Jams, marmalades and honey	1397
01.1.8.3.0	Chocolate	1655
01.1.8.4.0	Confectionery produts	1829
01.1.8.5.0	Ice creams	1537
01.1.9.1.0	Sauces and condiments	2550
01.1.9.2.0	Salt, spices and culinary herbs	2182
01.1.9.3.0	Baby food	571
01.1.9.4.0	Ready-made meals	2820
01.1.9.9.0	Yeasts and other preparates	1851

01.2.1.1.0	Coffee	1076
01.2.1.2.0	Tea	485
01.2.1.3.0	Cocoa and chocolate	96
01.2.2.1.0	Mineral waters	887
01.2.2.2.1	Soft drinks	835
01.2.2.2.2	Other soft drinks	448
01.2.2.3.0	Fruit and vegetable juices	1396
02.1.1.1.2	Spirits	1038
02.1.1.2.0	Alcoholic aperitifs	127
02.1.2.1.1	Table wines	2373
02.1.2.1.2	Quality wines	4032
02.1.2.1.3	Sparkling wines	819
02.1.2.3.1	Fortified wines	143
02.1.3.1.0	Lager beers	1035
05.6.1.1.1	Cleanings and maintenance products	1587
05.6.1.1.2	Dishwashing and detergents	425
05.6.1.2.0	Other non-durable small household products	3962
09.3.4.2.1	Pets and related produts	2247
09.3.4.2.2	Other pets products	1086
12.1.3.1.0	Non-electric appliances	926
12.1.3.2.1	Hair products	2546
12.1.3.2.1	Articles for personal hygiene and wellness	953
12.1.3.2.2	Body products	2465
12.1.3.2.3	Hygienic products	1905

Following the methodology illustrated in the first paragraph, we firstly run a CPD model for each available BH using weighted turnover.

As an example, in Figures 2.b.1 - 2.b.3 we report the results for Coffee, Fresh Pasta and Eggs BHs.

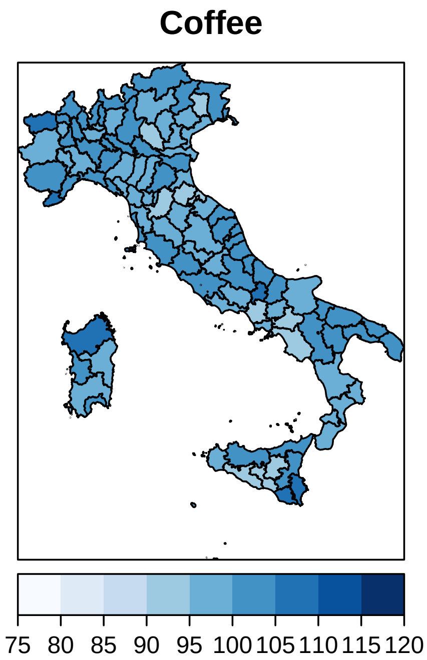


Figure 2.b.1.: PPPs at provincial level for Coffee BH

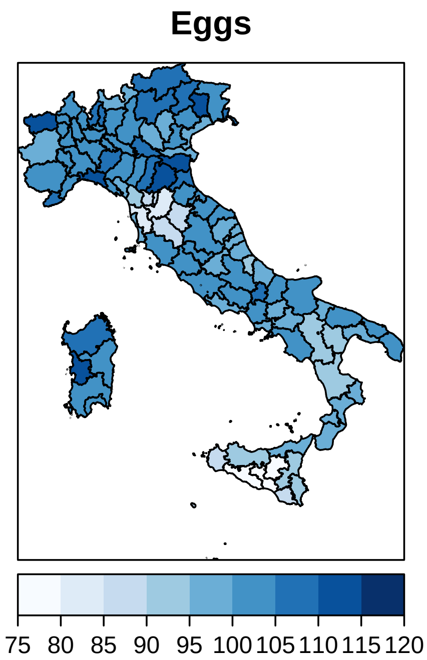


Figure 2.b.2.: PPPs at provincial level for Eggs BH

Table 2.b.2.: Descriptive statistics based on provincial PPPs

	Coffee	Fresh Pasta	Eggs
Min	92.69	83.03	75.90
Max	107.24	112.14	114.55
Mean	99.94	99.97	99.75
Std. Dev	3.32	6.45	7.38
CV	3.32	6.45	7.40

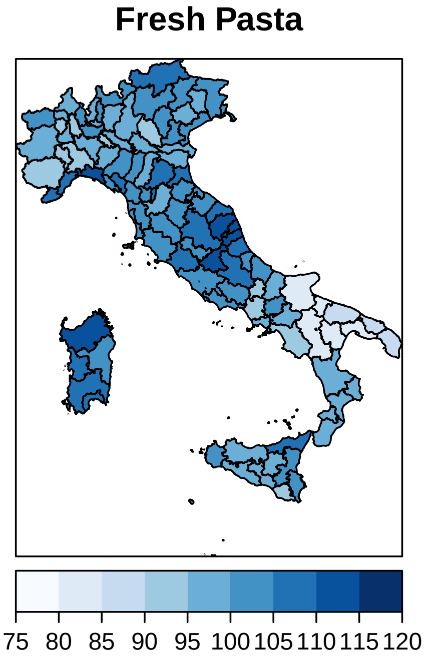


Figure 2.b.3.: PPPs at provincial level for Fresh Pasta BH

The North-South dualism is confirmed only for some BHs. In the case of Fresh Pasta, for which the coefficient of variation is equal to 6.45%, PPPs in the Northern provinces are generally higher than those in the Southern Italy. As illustrated in Figure 2.b.3, the less expensive provinces are Matera (83.03), Potenza (83.31) located in the Basilicata region and Foggia (83.67) located in Puglia region, while the most expensive provinces are Genova (112.14) located in Liguria region, Ascoli Piceno (111.06) and Fermo (111.02) located in Marche region. On the contrary, as shown in Figure 2.b.1, homogeneity in PPPs are observed for the Coffee BH; in this case the coefficient of variation across Italian provinces is equal to 3.32. Interesting results are provided for the Eggs BH, for which a high level of heterogeneity across Italian provinces is observed.

Using the results obtained in the first step, we computed the PPPs for two aggregates: “Food” and ”Non-food” products using as a weight the weighted turnover for each BH.

Results are reported in Table 2.b.4 and Figure 2.b.4.

Table 2.b.3.: PPPs for food and non-food aggregates at provincial level (Italy=100)

Provinces		PPPs food	PPPs non-food
North west			
Alessandria	AL	96.49	95.28
Aosta	AO	107.83	110.99
Asti	AT	94.76	92.02
Bergamo	BG	98.63	95.69
Biella	BI	98	95.7
Brescia	BS	100.22	98.13
Cuneo	CN	97.35	95.96
Como	CO	99.92	98.35
Cremona	CR	100.88	101.89
Genova	GE	107.5	108.12
Imperia	IM	104.1	106.53
Lecco	LC	98.44	96.74
Lodi	LO	100.74	99.42
Monza e della Brianza	MB	98.95	97.81
Milano	MI	99.54	97.92
Mantova	MN	100.23	99.96
Novara	NO	100.23	98.44
Pavia	PV	101.43	101.22
Sondrio	SO	99.06	97.15
La Spezia	SP	99.35	98.73
Savona	SV	104.13	105.68
Torino	TO	98.79	98.96
Varese	VA	98.79	97.49
Verbano-Cusio-Ossola	VB	102.07	101.08
Vercelli	VC	102.43	102.75
North west			
Belluno	BL	101.22	101.01
Bologna	BO	101.68	102.83
Bolzano	BZ	101.91	102.61
Forli'-Cesena	FC	96.71	95.24
Ferrara	FE	100.24	97.69
Gorizia	GO	102.36	101.47
Modena	MO	95.93	97.71
Piacenza	PC	100.04	99.13
Padova	PD	98.11	93.88

Pordenone	PN	96.83	95.09
Parma	PR	99.16	99.32
Ravenna	RA	100.38	97.98
Reggio nell'Emilia	RE	100.44	102.25
Rimini	RN	96.73	93.12
Rovigo	RO	98.54	94.44
Trento	TN	101.99	104.47
Trieste	TS	101.95	102.4
Treviso	TV	96.08	94.71
Udine	UD	102.68	100.76
Venezia	VE	98.39	96.52
Vicenza	VI	97.69	95.96
Verona	VR	93.75	92.84
Centre			
Ancona	AN	103.25	103.99
Ascoli Piceno	AP	105.14	106.55
Arezzo	AR	94.73	89.72
Firenze	FI	92.47	88.8
Fermo	FM	104.85	106.55
Frosinone	FR	101	103.04
Grosseto	GR	98.92	100.48
Livorno	LI	98.94	99.55
Latina	LT	100.05	99.62
Lucca	LU	96.7	93.38
Macerata	MC	103.7	105.16
Massa Carrara	MS	98.56	96.69
Perugia	PG	100.38	102.14
Pisa	PI	93.57	90.3
Prato	PO	94.7	88.22
Pistoia	PT	95.3	89.7
Pesaro e Urbino	PU	97.93	96.58
Rieti	RI	101.35	103.65
Roma	RM	103.15	103.09
Siena	SI	97.5	94.36
Terni	TR	101.27	103.05
Viterbo	VT	102.83	106.38
South			
L'Aquila	AQ	105.4	108.2
Avellino	AV	98.74	100.78
Bari	BA	97.96	100.39
Benevento	BN	101.41	104.73
Brindisi	BR	98.05	100.34
Barletta-Andria-Trani	BT	98.76	100.9

Campobasso	CB	100.91	102.01
Caserta	CE	96.14	96.52
Chieti	CH	102.86	103.21
Cosenza	CS	100.07	102.42
Catanzaro	CZ	98.27	100.8
Foggia	FG	98.1	100.28
Isernia	IS	103.78	105.02
Crotone	KR	103.78	102.42
Lecce	LE	99.96	101.55
Matera	MT	98.56	100.39
Napoli	NA	97.6	99
Pescara	PE	102.2	102.28
Potenza	PZ	102.83	105.55
Reggio di Calabria	RC	101.44	103.63
Salerno	SA	98.14	99.17
Taranto	TA	97.69	100.45
Teramo	TE	103.42	103.38
Vibo Valentia	VV	102.83	101.23
Islands			
Agrigento	AG	97.28	100.78
Cagliari	CA	100.14	100.81
Carbonia-Iglesias	CI	102.86	100.81
Caltanisetta	CL	102.86	106.87
Catania	CT	102.58	106.87
Enna	EN	98.27	100.78
Messina	ME	105.04	107.14
Nuoro	NU	101.55	102.02
Ogliastra	OG	101.55	100.81
Oristano	OR	105.18	107.89
Olbia-Tempio	OT	105.18	100.81
Palermo	PA	99.18	102.79
Ragusa	RG	104.21	103.71
Siracusa	SR	104.09	108.75
Sassari	SS	101.9	100.87
Trapani	TP	99.49	102.3
Medio Campidano	VS	102.83	100.81

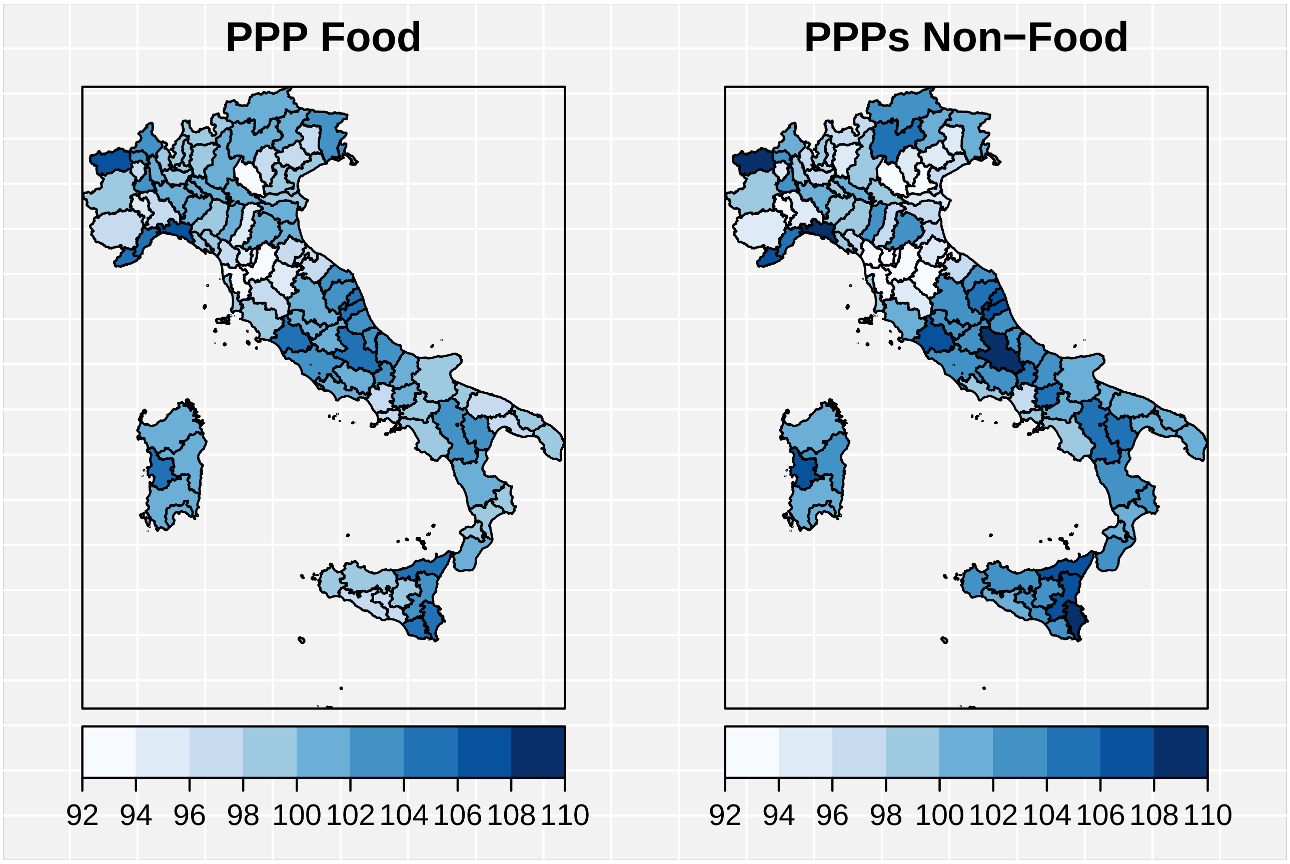


Figure 2.b.4.: PPPs at provincial level for food and non-food aggregates

From Table 2.b.4 and Figure 2.b.4 we can observe a high level of price heterogeneity across Italian Provinces both for food and non-food aggregates. For food products the most expensive provinces are Aosta and Genova (with PPPs equal to 107.83 and 107.5 respectively) located in the Northern Italy, while the less expensive provinces are Florence and Pisa (with PPPs equal to 92.47 and 93.57 respectively) located in the Central Italy. For non-food products, the most expensive provinces are Aosta and L'Aquila (with PPPs equal to 110.99 and 108.20 respectively), while the less expensive provinces are Prato and Pistoia (with PPPs equal to 88.22 and 89.70 respectively) located in the Tuscany region, in Central Italy.

2.b.3.4. PPPs for the first quintile of the price distribution at regional level

In order to calculate PPPs for the lowest price in the price distribution, in particular for the first quintile of the price distribution, we used data for the 20 regions (NUTS 2 level), by applying a two-step procedure for the food BHs. Results are reported in Figure 2.b.5. In this case, we provide PPPs at regional level, since at Provicial level there was not a reliable level of overlap. Results in Table Figure 2.b.5 provede PPPs estimates for food aggregates.

Table 2.b.4.: PPPs for food aggregates for first quintile and total distribution of price (Italy=100)

REGIONS	PPPs FIRST	PPPS TOTAL

	QUINTILE	DISTRIBUTION
North-west		
Liguria	101.78	102.14
Lombardia	100.04	99.28
Piemonte	99.34	98.75
Valle d'Aosta	107.39	105.86
North-east		
Emilia-Romagna	99.10	99.08
Friuli Venezia-Giulia	100.96	100.28
Trentino Alto Adige	100.78	100.20
Veneto	97.82	96.94
Centre		
Lazio	100.56	101.04
Marche	100.97	101.11
Toscana	96.23	95.40
Umbria	100.02	98.90
South		
Abruzzo	100.95	101.96
Basilicata	100.43	101.79
Calabria	98.32	100.26
Campania	96.90	98.00
Molise	101.29	99.73
Puglia	97.70	97.53
Islands		
Sardegna	100.17	100.34
Sicilia	99.74	101.90

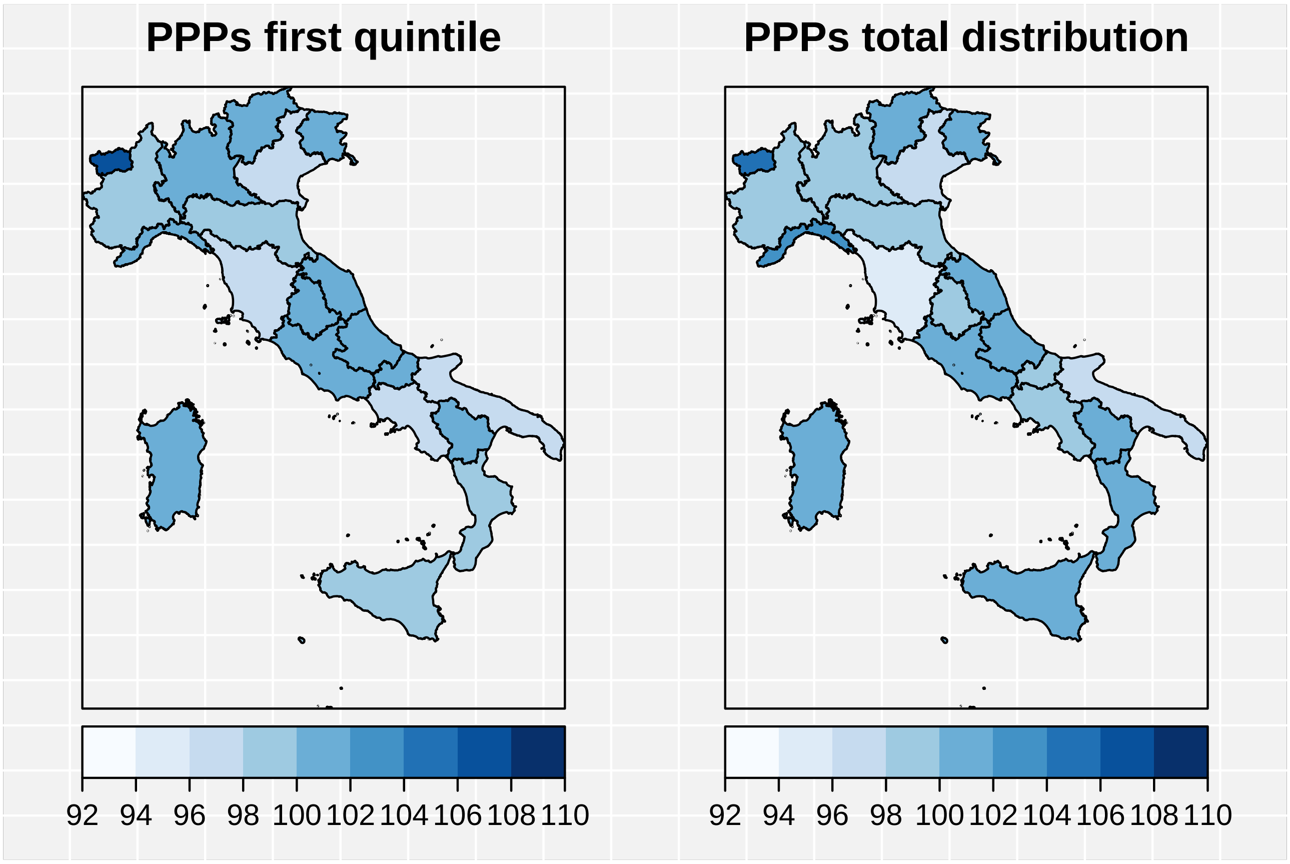
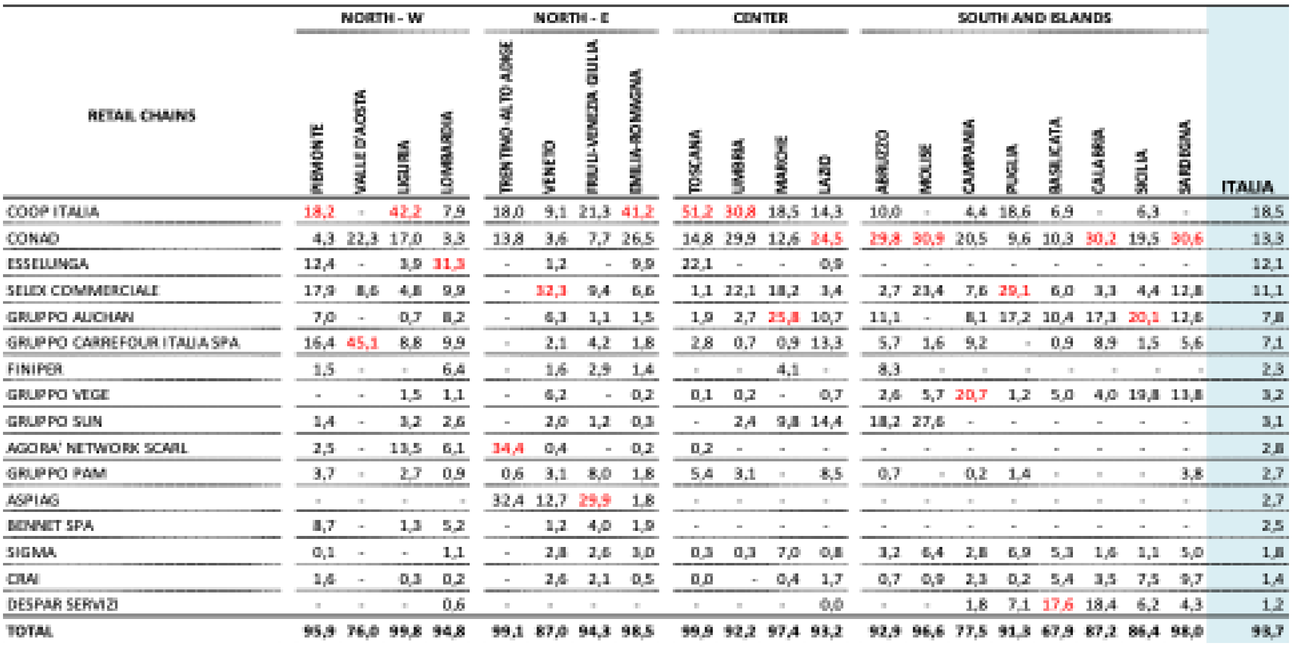


Figure 2.b.5.: PPPs at regional level for first quintile and total distribution of the price

As reported in Figure 2.b.5, for food aggregates calculated for the first quintile of the distribution, the most expensive regions are Valle d'Aosta and Liguria (with PPPs equal to 107.39 and 101.78 respectively), while the less expensive regions are Toscana and Campania (with PPPs equal to 96.23 and 96.90 respectively). Similar results are obtained if we considered the total distribution of the price. Indeed the most expensive regions are Valle d'Aosta and Liguria (with PPPs equal to 105.86 and 102.14 respectively), while the less expensive regions are Toscana and Veneto (with PPPs equal to 95.40 and 96.94 respectively).

As pointed out in the section 2.b.2, there are various advantages in using scanner data to compute SPIs and our PPPs. On the other hand, the like to like approach may have some limitations. Indeed, in order to use strictly comparable products, it is possible that some products are excluded, since they are produced and consumed at local level. In our application we had to exclude some BHs due to the insufficient overlap across provinces (i.e. whole milk, low-fat milk, olive oil, aged cheese, other cheese, lager beer and frozen seafood). It is also worth noting that PPP results may be influenced by the characteristics of the modern retail trade which is not uniformly distributed across Italian territory in terms of types of retail chains and market share. From Figure 2.b.6 it is clear that in some Southern regions the share covered by the retail chains is lower than that observed in the North of Italy. In addition, consumer choice among the different distributional channel may be considered. In Southern regions consumers tend to buy food and non- food products in open markets and traditional shops more frequently than consumers in Northern regions.

Figure 2.b.6.: Scanner data: % market shares (hypermarket + supermarket) – year 2016



	NORTH - W				NORTH - E				CENTER				SOUTH AND ISLANDS								
RETAIL CHAINS	PIEMONTE	VALLE D'AOSTA	LIGURIA	LOMBARDIA	TRENTINO ALTO ADIGE	VENETO	FRIULI-VENEZIA GIULIA	EMILIA-ROMAGNA	TOSCANA	UMBRIA	MARCHE	LAZIO	ABRUZZO	MOLISE	CAMPANIA	PUGLIA	BASILICATA	CALABRIA	SICILIA	SARDEGNA	ITALIA
COOP ITALIA	18,2	-	42,2	7,9	18,0	9,1	21,3	41,2	51,2	30,8	18,5	14,3	10,0	-	4,4	18,6	6,9	-	6,3	-	18,5
CONAD	4,3	22,3	17,0	3,3	13,8	3,6	7,7	26,5	14,8	29,9	12,6	24,5	29,8	30,9	20,5	9,6	10,3	30,2	19,5	30,6	13,3
ESSELUNGA	12,4	-	3,9	31,3	-	1,2	-	9,9	22,1	-	-	0,9	-	-	-	-	-	-	-	-	12,1
SELEX COMMERCIALE	17,9	8,6	4,8	9,9	-	32,3	9,4	6,6	1,1	22,1	18,2	3,4	2,7	23,4	7,6	29,1	6,0	3,3	4,4	12,8	11,1
GRUPPO AUCHAN	7,0	-	0,7	8,2	-	6,3	1,1	1,5	1,9	2,7	25,8	10,7	11,1	-	8,1	17,2	10,4	17,3	20,1	12,6	7,8
GRUPPO CARREFOUR ITALIA SPA	16,4	45,1	8,8	9,9	-	2,1	4,2	1,8	2,8	0,7	0,9	13,3	5,7	1,6	9,2	-	0,9	8,9	1,5	5,6	7,1
FINIPER	1,5	-	-	6,4	-	1,6	2,9	1,4	-	-	4,1	-	8,3	-	-	-	-	-	-	-	2,3
GRUPPO VEGE	-	-	1,5	1,1	-	6,2	-	0,2	0,1	0,2	-	0,7	2,6	5,7	20,7	1,2	5,0	4,0	19,8	13,8	3,2
GRUPPO SUN	1,4	-	3,2	2,6	-	2,0	1,2	0,3	-	2,4	9,8	14,4	18,2	27,6	-	-	-	-	-	-	3,1
AGORA' NETWORK SCARL	2,5	-	13,5	6,1	34,4	0,4	-	0,2	0,2	-	-	-	-	-	-	-	-	-	-	-	2,8
GRUPPO PAM	3,7	-	2,7	0,9	0,6	3,1	8,0	1,8	5,4	3,1	-	8,5	0,7	-	0,2	1,4	-	-	-	3,8	2,7
ASPIAG	-	-	-	-	32,4	12,7	29,9	1,8	-	-	-	-	-	-	-	-	-	-	-	-	2,7
BENNET SPA	8,7	-	1,3	5,2	-	1,2	4,0	1,9	-	-	-	-	-	-	-	-	-	-	-	-	2,5
SIGMA	0,1	-	-	1,1	-	2,8	2,6	3,0	0,3	0,3	7,0	0,8	3,2	6,4	2,8	6,9	5,3	1,6	1,1	5,0	1,8
CRAI	1,6	-	0,3	0,2	-	2,6	2,1	0,5	0,0	-	0,4	1,7	0,7	0,9	2,3	0,2	5,4	3,5	7,5	9,7	1,4
DESPAR SERVIZI	-	-	-	0,6	-	-	-	-	-	-	-	0,0	-	-	1,8	7,1	17,6	18,4	6,2	4,3	1,2
TOTAL	95,9	76,0	99,8	94,8	99,1	87,0	94,3	98,5	99,9	92,2	97,4	93,2	92,9	96,6	77,5	91,8	67,9	87,2	86,4	98,0	98,7

2.b.4. Spatial price indexes: ASESD approach

2.b.4.1. Estimation of Spatial Consumer Price Indexes for the Italian Provinces

The results we present in this section concern the computation of spatial consumer price indices (SN-SCPI) for Italian provinces by using the scanner data of the products sold in modern distribution chains referring to the year 2018 and only to the products (barcodes or GTINs) in food and beverages categories, excluding fresh food. Usually the information on products' quantities is reported in terms of grams and milliliter, but sometimes in units; given that we needed to use comparable prices, we discarded about 17,000 quotations expressed in units.

To estimate the SN-SCP for each of the 103 provinces, two-step procedure has been followed.

In the first step, we computed the average unit price at level of province by considering the unit value prices from the consumer side (or points of view). In applying the principle of comparability, we did not follow a very tight way by considering the comparisons of the 'like to like' items (products). Instead, we applied the principle at a different level, the level products' groups, and exactly at the level of the 102 groups of the ECOICOP-8-digit classification. The hypothesis is that the elementary products (items) within a group are sufficiently similar, so that consumers are generally indifferent about the choice of the product that in any case satisfy the same consumer needs (may be giving him/her the same utility), even if the brand, quality, etc. is different. The comparison is therefore done by considering the average level of prices of the group of products purchased in the different provinces, considering the basket of elementary products that the consumers of each province have really purchased[2].

In what follows we define the weighted mean price $\bar{p}_{ij}$ for ECOICOP-8-digit j and province i. Let r_{ijk} and q_{ijk} be the annual turnover and the total quantity sold respectively of item k belonging to ECOICOP-8-digit j in province i. These quantities are estimated by Istat using the scanner data and the sampling weights computed according to the survey design summarised in section 2.b.2. Let u_{ijk} be the quantity of the item ijk in terms of gr. or ml. For each item we define its annual price per gr. or ml. as

$$p_{ijk} = \frac{\frac{r_{ijk}}{q_{ijk}}}{u_{ijk}}.$$

Then, for each item we define its relative weights in term of turnover as

$$w_{ijk} = \frac{r_{ijk}}{\sum_{k=1}^{n_{ij}} r_{ijk}},$$

where n_j is the number of items in the jth ECOICOP-8-digit aggregation and the ith province. Finally, the weighted mean price is:

$$\bar{p}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} p_{ijk} w_{ijk}.$$

Therefore, $\bar{p}_{ij}$ is the weighted mean price per gr. or ml. for products in ECOICOP-8-digit j and province i.

The second step is devoted to the aggregation of 102 average level of prices to estimate the provincial Sn-SCPI. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces.

To compute the SPIs at provincial level, we adapt a Country Product Dummy model (Laureti and Rao, 2018). The products are aggregated by province and ECOICOP-8-digit classification, for a total of 103 provinces and 102 ECOICOP-8-digit. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces. The CPD model we propose is as follows:

$$\log \bar{p}_{ij} = \alpha_0 + \alpha_i D_i + \beta_j I_j + \varepsilon_{ij}, \quad i = 1, \dots, 103 \quad j = 1, \dots, 102, \qquad (2.b.8)$$

where D_i is a vector equal 1 if the mean price is in province i and 0 otherwise, I_j is equal 1 if the mean price belongs to jth ECOICOP-8-digits and 0 otherwise. The index i is for the provinces and the index j is for the ECOICOP-8-digit. The error is $\varepsilon_{ij} \sim N(0, \sigma^2)$.

To take into account the different level of the turnover between the ECOICOP-8-digit aggregates, we estimate the model (2.b.8) using weighted least squares, where the weights are computed as

$$wls_{ij} = \frac{\sum_{k=1}^{n_{ij}} r_{ijk}}{\sum_{k=1}^{n_i} r_{ijk}},$$

[2] The value of the average level of prices of the different provinces could be affected by the different typologies of families (number of components, age, etc.) in the provinces (Istat, 2009, Biggeri and Laureti, 2018). To obtain more precise comparison among the different averages, it could be necessary to make some standardization of the provincial averages. This is an issue that the unit of research will deepen in a near future

that is the ratio between the total turnover of one aggregate in one province, and the total turnover in the province (n_i is the number of items in the ith province).

Model (2.b.8) – as it is specified – is not identified, because the D_is vectors are a linear combination of the constant. Therefore we impose the constraint $\alpha_1 = 0$ so that the model is identified. Once the model is estimated, from the data we obtain the estimates of the SPIs at provincial level by $\exp(\hat{\alpha}_i)$, where $\hat{\alpha}_i$ is the estimate of α_i. The coefficient α_i is the difference of fixed effects connected with the province i compared with the base province $i = 1$. To use as a reference Italy instead of area 1, the coefficient $\hat{\alpha}_i$ has been adjusted following Suits (1984). In this way, α_i represents the fixed effect of province i compared to Italy. Thus, the quantity $\exp(\hat{\alpha}_i)$ represents the spatial price index for food in province i with respect to Italy, and it is also called purchasing power parity of province i (PPP_i).

An advantage of the use of CPD models is that we can obtain p-values for the estimated coefficients. Following Suits (1984) we derive the p-values for the rescaled $\hat{\alpha}_i$s which are not reported here. Setting a I type error equal to 0.1 we observed 43 provinces for which we don't reject the hypothesis $\alpha_i = 0$ which correspond to a SPI equal 1. Out of these 43 provinces 17 are located in the north, 18 in the center, and 8 in the south of Italy.

The SPIs estimated at the province level can be used for many purposes. One of these purposes is to adjust the national poverty line at the province level, by this way relative poverty estimates take into account the different purchase power within the country. An application to Italian data is shown later in the Deliverable. The SPIs estimated according to model (2.b.8) are based on mean prices of specific headings (ECOICOP-8-digit), therefore the adjustment of the national poverty line is not poor specific. As an alternative, our method can be easily extended to produce SPIs related to specific quantiles of the distribution of the prices of specific headings. By this way we can adjust the national poverty line using SPIs based on lower prices instead of mean prices, which are reasonably related to poor households. To obtain such SPIs the model (2.b.8) is modified as follows:

$$\log Q(\tau, p)_{ij} = \gamma_0 + \gamma_i D_i + \beta_j I_j + \varepsilon_{ij}, \quad i = 1, \dots, 103 \quad j = 1, \dots, 102, \tag{2.b.9}$$

where $Q(\tau, p)_{ij}$ is the quantile of order τ of the unit prices (p_{ijk}) belonging to heading j (ECOICOP-8-digit) and province i. To estimate γ_i we use the same method used to estimate α_i in model (2.b.8).

For example, setting $\tau = 0.2$ we can obtain the estimates of spatial price indices related to the cheaper prices for each Italian provinces, which we denote as SPI($\mathrm{Q}_{0.2}$)'s and shown in 2.b.5.

2.b.4.2. Results of the estimation of SPIs and SPI($\mathbf{Q}_{0.2}$)'s

The estimates of the SPIs and SPI($\mathrm{Q}_{0.2}$)'s we obtained are shown in table 2.b.5. Moreover, we show a choropleth map of estimated SPIs (based on model (2.b.8)) for the Italian provinces in figure 2.b.7 on the left.

Table 2.b.5.: Estimates of SPI's based on the mean prices and on quantile 0.2.

Prov.	SPI(Mean)	SPI($Q_{0.2}$)	Region
TO	1.123	1.043	PIEMONTE
VC	1.044	1.058	PIEMONTE
NO	1.102	1.094	PIEMONTE
CN	1.035	1.012	PIEMONTE
AT	0.914	0.960	PIEMONTE
AL	1.087	1.057	PIEMONTE
BI	0.973	1.021	PIEMONTE
VB	1.108	1.134	PIEMONTE
AO	1.047	1.143	VALLE D'AOSTA
IM	1.057	1.084	LIGURIA
SV	1.072	1.044	LIGURIA
GE	1.069	1.021	LIGURIA
SP	1.087	1.090	LIGURIA
VA	1.100	1.032	LOMBARDIA
CO	1.134	1.106	LOMBARDIA
SO	0.996	0.938	LOMBARDIA
MI	1.098	1.027	LOMBARDIA
BG	1.110	1.055	LOMBARDIA
BS	1.097	1.047	LOMBARDIA
PV	1.132	1.101	LOMBARDIA
CR	1.105	1.074	LOMBARDIA
MN	1.035	1.035	LOMBARDIA
LC	1.100	1.078	LOMBARDIA
LO	1.005	0.992	LOMBARDIA
MB	1.119	1.086	LOMBARDIA
BZ	0.974	0.970	TRENTINO-ALTO ADIGE
TN	1.023	0.964	TRENTINO-ALTO ADIGE
VR	1.024	0.979	VENETO
VI	1.017	1.000	VENETO
BL	0.930	0.956	VENETO
TV	1.039	1.016	VENETO
VE	1.051	1.010	VENETO
PD	1.010	0.985	VENETO
RO	1.000	1.012	VENETO
UD	1.077	1.062	FRIULI-VENEZIA GIULIA
GO	0.998	1.003	FRIULI-VENEZIA GIULIA
TS	1.012	1.018	FRIULI-VENEZIA GIULIA
PN	0.972	0.958	FRIULI-VENEZIA GIULIA
PC	1.042	1.058	EMILIA-ROMAGNA
PR	1.064	1.038	EMILIA-ROMAGNA

RE	1.064	1.006	EMILIA-ROMAGNA
MO	1.060	1.002	EMILIA-ROMAGNA
BO	1.085	1.055	EMILIA-ROMAGNA
FE	1.026	1.043	EMILIA-ROMAGNA
RA	0.985	0.927	EMILIA-ROMAGNA
FC	1.025	0.956	EMILIA-ROMAGNA
RN	1.008	0.971	EMILIA-ROMAGNA
PU	0.981	0.952	MARCHE
AN	1.075	1.068	MARCHE
MC	1.032	1.033	MARCHE
AP	1.000	1.020	MARCHE
FM	0.997	1.045	MARCHE
MS	0.984	1.015	TOSCANA
LU	1.028	0.999	TOSCANA
PT	0.938	0.925	TOSCANA
FI	1.005	0.939	TOSCANA
LI	1.040	0.980	TOSCANA
PI	1.028	0.975	TOSCANA
AR	0.974	0.975	TOSCANA
SI	0.974	0.991	TOSCANA
GR	1.010	0.979	TOSCANA
PO	1.000	1.034	TOSCANA
PG	0.999	0.958	UMBRIA
TR	0.937	0.949	UMBRIA
VT	1.024	1.000	LAZIO
RI	0.982	0.983	LAZIO
RM	1.062	0.986	LAZIO
LT	1.024	1.013	LAZIO
FR	1.016	1.018	LAZIO
CE	0.931	0.946	CAMPANIA
BN	0.828	0.865	CAMPANIA
NA	0.964	0.963	CAMPANIA
AV	0.898	0.913	CAMPANIA
SA	0.904	0.900	CAMPANIA
AQ	1.115	1.125	ABRUZZO
TE	1.022	1.032	ABRUZZO
PE	1.012	1.026	ABRUZZO
CH	1.061	1.033	ABRUZZO
CB	0.989	1.022	MOLISE
IS	0.957	1.068	MOLISE
FG	0.940	0.946	PUGLIA
BA	0.974	0.959	PUGLIA
TA	0.947	0.954	PUGLIA

BR	0.929	0.943	PUGLIA
LE	0.915	0.909	PUGLIA
BT	0.884	0.888	PUGLIA
PZ	0.827	0.905	BASILICATA
CS	0.908	0.913	CALABRIA
CZ	0.893	0.908	CALABRIA
RC	0.882	0.937	CALABRIA
VV	0.845	0.956	CALABRIA
TP	0.889	0.920	SICILIA
PA	0.928	0.968	SICILIA
ME	0.956	0.981	SICILIA
AG	0.707	0.820	SICILIA
CT	0.975	0.995	SICILIA
RG	0.915	0.943	SICILIA
SR	0.931	0.995	SICILIA
SS	1.062	1.091	SARDEGNA
NU	0.927	0.972	SARDEGNA
CA	1.041	1.081	SARDEGNA
OR	0.978	1.081	SARDEGNA
SU	1.024	1.077	SARDEGNA

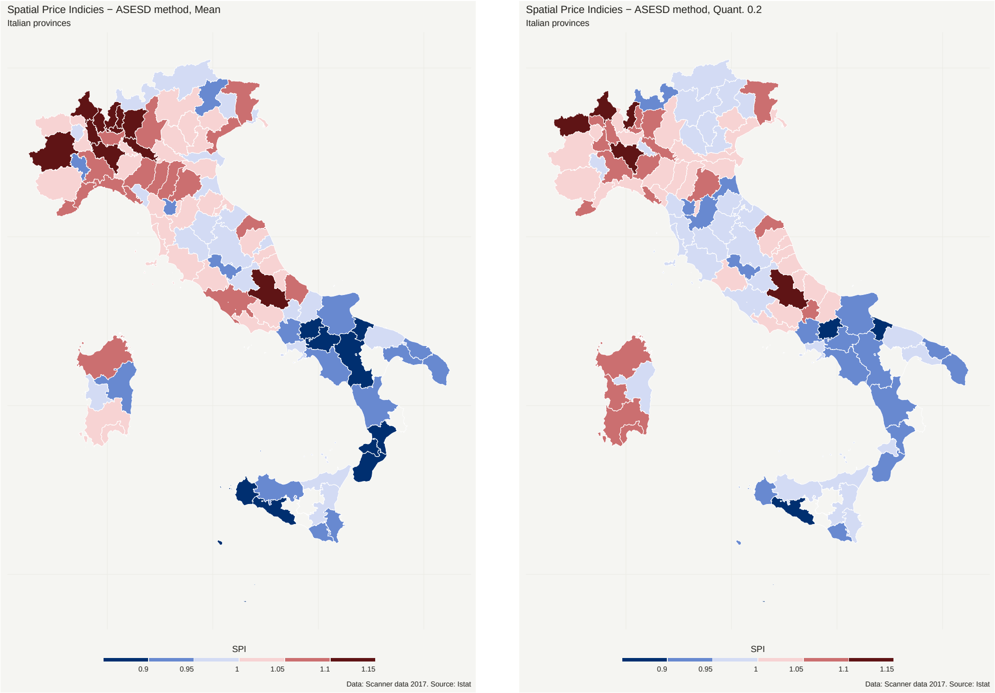

Spatial Price Indicies – ASESD method, Mean
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.15
Data: Scanner data 2017. Source: Istat
Spatial Price Indicies – ASESD method, Quant. 0.2
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.15
Data: Scanner data 2017. Source: Istat

Figure 2.b.7.: Choropleth map of SPI computed according to ASESD method. SPI obtained using mean unit prices (left) and quantile 0.2 of unit prices (right).

The results we obtained are somehow expected. Indeed, provinces in the south of Italy show SPI smaller than 1, while provinces in the north show values greater than 1. However, there are exceptions, provinces in the north-east Alps mountains show SPI below 1, even if they are close. Provinces in the center of Italy have SPIs close to 1, with some evidence of SPI lower than 1 for provinces located in the Appennino mountains (middle of central Italy), and SPI greater than 1 for the provinces located on the seaside, both Adriatic (east), Ligure and Tirreno (west). The lowest SPI is estimated for the province of Agrigento (AG), in Sicily (south of Italy), while the highest is in the province of Como (CO), in Lombardia region (north of Italy). The provinces with the highest SPI are all located in the north-west, but Aquila (AQ) located in Abruzzo, a region in the south[3] of Italy.

As it concerns the estimates of SPI($Q_{0.2}$) mapped in figure 2.b.7 on the right, we recall that they are obtained modifying model (2.b.8). In the results we don't reject the null hypothesis that $\gamma_i = 0$, that is SPI($Q_{0.2}$)=1, for 13 provinces in the north, 15 in the center and 5 in the south of Italy. Looking at the results of the estimation of the SPI($Q_{0.2}$)'s we observe that many provinces in the north-west and on the Adriatic seaside, excluding Puglia provinces, show a SPI($Q_{0.2}$) greater than 1, while many provinces in the north-east, center and south of Italy show SPI($Q_{0.2}$) smaller than 1. Sardegna provinces show SPI($Q_{0.2}$) greater than 1, except for Nuoro.

When we build spatial price indices using lowest prices we observe a similar behaviour of indices built with mean prices, however there are differences, for example the province of Rome has SPI($Q_{0.2}$) = 0.986 (pval = 0.4) and SPI = 1.06 (pval = 0.007), Isernia has SPI($Q_{0.2}$) = 1.07 (pval = <0.0001) and SPI = 0.957 (pval = 0.046). Other 19 provinces show discordance between point estimates of SPI($Q_{0.2}$) and SPI, 3 in the south and 16 in north and central Italy.

2.b.5. A general analysis of the results of the experiments: some concluding remarks

The results obtained with the two experiments are undoubtedly interesting. Actually the estimations of the PPPs and SPIs (according the ASEDS approach) at provincial level are quite different as we see from Figure 8 that reports the indexes computed with the two approaches using the same scale.

As illustrated in the previous sections, the PPPs obtained with the WB method are smoother and indicate that the prices are lower in the provinces of Tuscany, located in Central Italy, and in some northern provinces. However, also some provinces in the south of the Country show values below 1. On the opposite, the SPIs computed using the ASESD methodology reported in the section, show a general north/south divide, although with some exceptions, as already commented in the previous section. Another main difference is that the range of the ASESD SPIs values is higher, as it covers the lower and the higher classes of values represented in the Figure (values lower than 0.9 and higher than 1.1).

The results of the two experiments are therefore different, but we have to take into account that the followed procedures are different too. As suggested in the World Bank book (2013), when we compute the within-country PPPs one would expect some internal consistency. Price level in poor areas should be generally lower than those in richer areas and showed a similar pattern across the basic headings. To check this hypothesis, we have done a comparison between the two computed indexes and the value

[3] Actually the Abruzzo region is in the center of the country, however, for historical reasons it is included among the southern regions

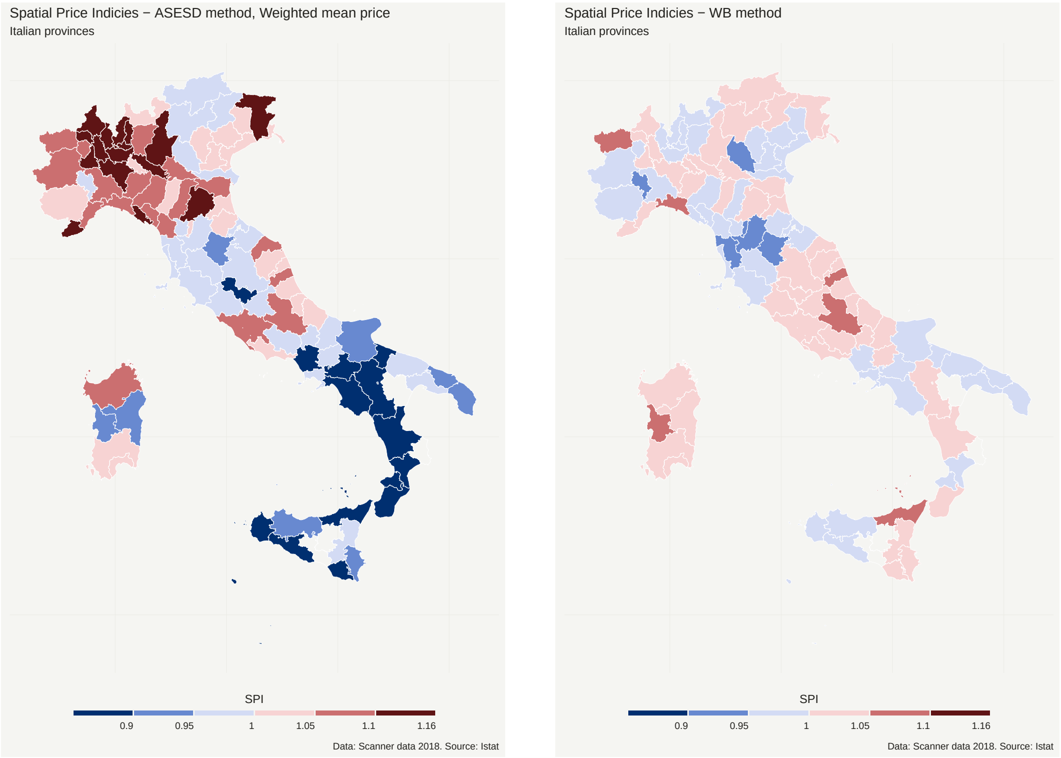

Spatial Price Indicies − ASESD method, Weighted mean price
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.16
Data: Scanner data 2018. Source: Istat
Spatial Price Indicies − WB method
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.16
Data: Scanner data 2018. Source: Istat

Figure 2.b.8.: Choropleth map of SPI computed according to ASESD and the WB method.

added per capita by Italian province as reported in the figure 2.b.9. It is evident that the Indexes computed with ASESD method satisfy in some way the previous mentioned consistency, while the PPPs do not satisfy it (in fact the correlation coefficients are about + 0,60 for the first index and -0,10 for the second.)

As explained in the previous subsections, the two sets of SPIs values have been computed with different methodologies and with slightly different sets of data. We think that the difference in the results mainly depends on the different methodologies that have been used: the WB approach is based on like-to-like product comparisons, and we expect that the range of the price for the same product cannot vary so highly in the hypermarkets and supermarkets, although located in different areas of the Italian territory. The influence of the political prices of the different commercial chains should be analyzed, and to have a more clear picture of the reasons of the differences, the analyses should be done at a disaggregated level. Unfortunately, a finer comparison at Basic heading level is not possible in this moment, but we plan to better investigate in the future. In any case the results obtained are interesting and useful from scientific and official statistics point of view. We think that the unit of research has to continue the experiments, as well as the same experiments should be conducted also by the units of research of other European countries.

Finally, we would like to inform the readers, that, outside of the computation of the SN-SCPIs by using scanner data, the unit of research has done the computation of sub-national SPIs for Housing costs which acquires particular importance since housing costs can be a substantial financial burden to households, especially for low-income families. Such latest indexes have surely an autonomous

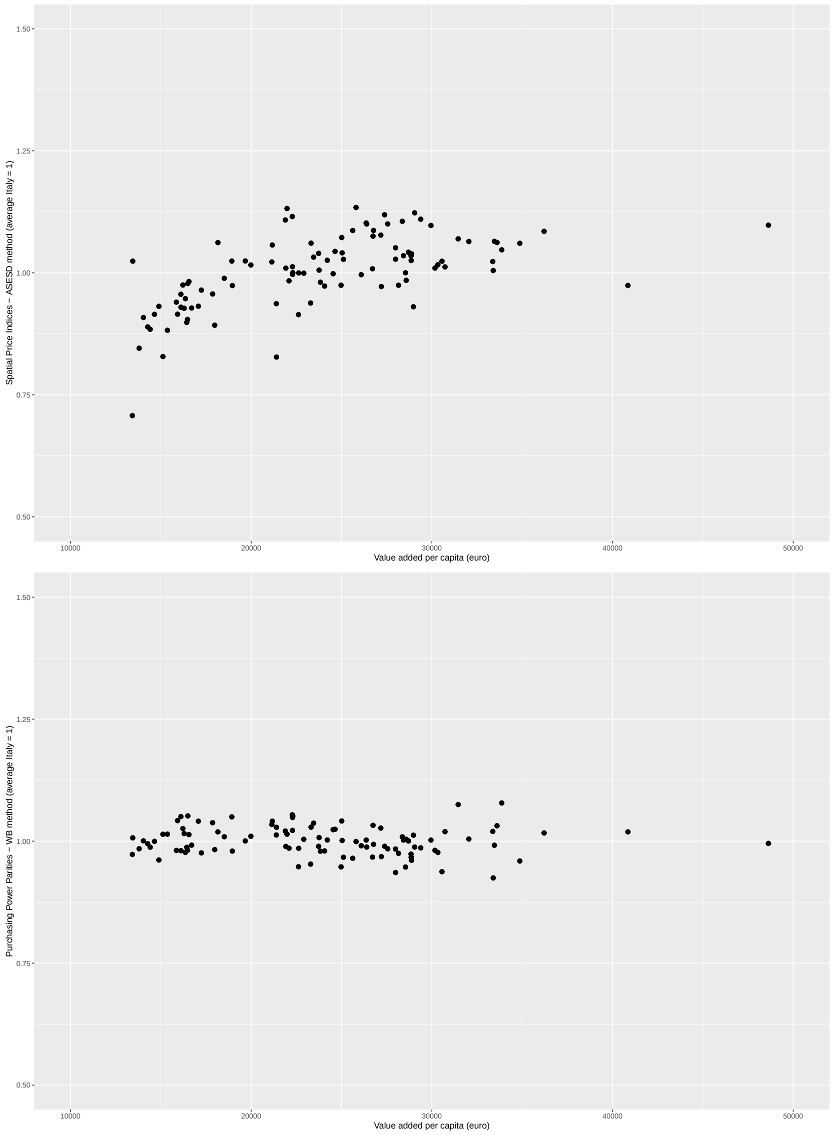


Figure 2.b.9.: Scatterplot of the SPis and PPPs versus the value added per capita, Italian provinces.

relevance in assessing and comparing poverty and in designing housing policies for the poor at a local level, but they can also be used as proxies of the general sub-national household consumption SPIs or in combination with them, for adjusting poverty thresholds. We computed the SPIs for Housing Rents (SPIHRs) by using two different sources of data coming from: (i) Archives on the evaluation of rents for all the 7,914 Italian municipalities, made by the Revenue and Tax Agency; (ii) Rents and house information collected through the Household Expenditure Survey (HES, with a sample of about 25,000 units). In both cases, the hedonic regression models, by including location and house characteristics, have been used to obtain the SPIHRs for the Italian regions. The results obtained, making the Italian average=100, show significant level differences across the various Italian regions: (i) by using the revenue and tax agency data for all type of dwellings the max is 152,73 and the min is 53,47 and for the economic apartments the max is 156,87 and the min 48,48; (ii) by using the HES data for the rented dwellings the max is 156,80 and the min 50,60. These results point out the importance and the possibility of calculating the SPIHR in Italy on a regular basis.

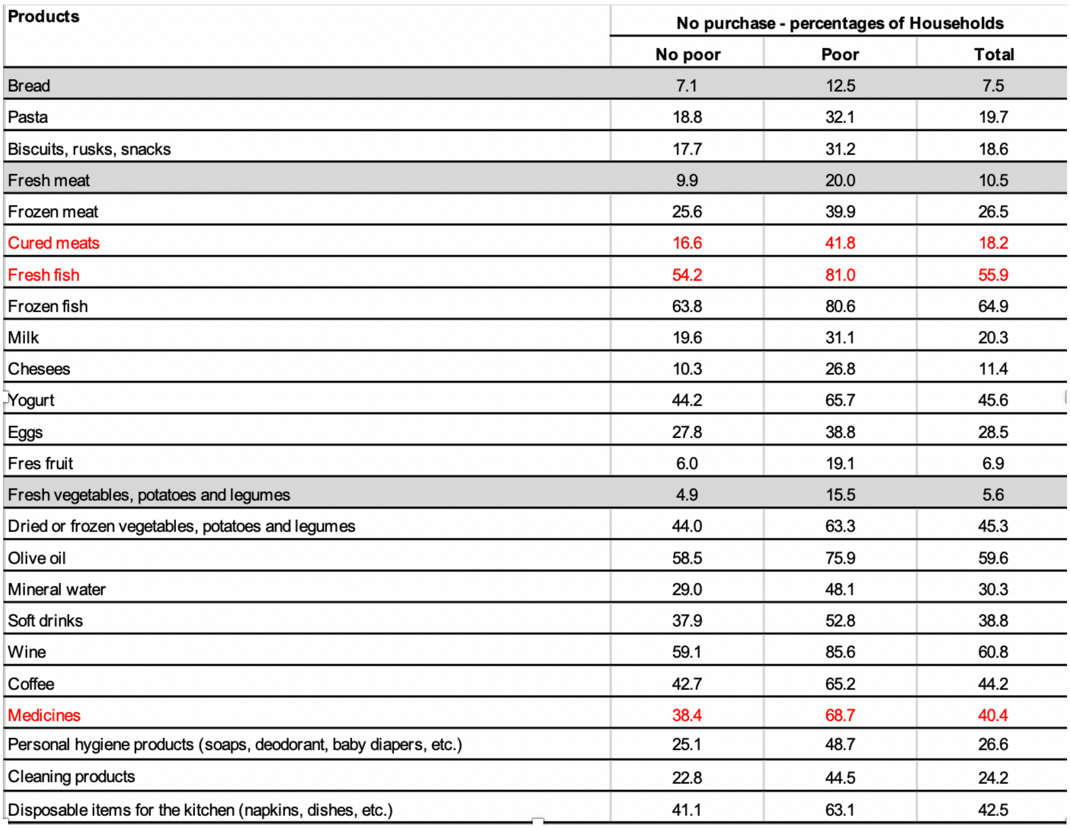
2.b.6. Where people in condition of absolute poverty purchase some large consumption products

In the two-weeks Diary (aimed at investigating the more frequent expenditures referred to large consumption products) of Italian Household Expenditure Survey, HES, since 2015 Istat asks households to indicate, in a dedicated section, the type of outlets where they have purchased a list of 25 products. The products are the most frequently purchased and the types of outlets listed are seven: traditional shop, open market and street vendors, hard discount, hypermarkets and supermarkets, department stores and outlet chains, farm or direct producer, internet. Households are asked to indicate the two places where they have more frequently purchased that specific good or to select the option 'not purchased'. On 2019 HES data an analysis has been carried out with the objective of detecting possible differences in terms of choices of types of outlets between households over (non poor) and households on or below the thresholds of absolute poverty (poor) calculated and updated each year by Istat. In terms of frequency of purchase, fresh vegetables and fruit, bread, meat and cheeses are the products purchased most frequently (figure 2.b.10) by the entire population: more than 85% buys at least once in two weeks. At the opposite end of this ranking we find toys and video games, frozen fish, wine and olive oil that are purchased by 40%, and even less, of the population considering the time span of two weeks and thus less frequently. If we limit the observation only to households in conditions of absolute poverty, differences emerge in purchasing attitudes. In particular, the purchase of some essential products such as bread, milk and eggs is (relatively) more important for poor families. On the contrary, the purchase of cured meats loses importance, (it is at the 6th place in the overall ranking of products purchased at least once and falls to 11th place in the case of absolute poor households).

Particular attention should be paid to the purchase of medicines whose general frequency is relatively low, about 60% of the population makes at least one purchase in the period under review (16th place in the ranking). In the case of the poor, the percentage drops to 40% (20th in the ranking). The reasons for this difference are probably mainly the use of drugs only in situations of particular gravity and a greater possibility of obtaining drugs for free from the National Health Service. In red, in figure 2.b.10, we find those products that show a wider difference between poor and non-poor in terms frequency of purchase. In addition to the case of cured meats and medicines, that of fresh fish is highlighted. The products with the smallest differences are highlighted in gray.

In addition to bread, milk and eggs already mentioned, there is a frequent consumption, even by the poor, of fresh meat and vegetables. With reference to the entire population, the most suitable place for shopping is, in all cases, the super-hypermarket with only two exceptions: medicines and toys (figure 2.b.11). The other two main shopping places are the same in almost all the cases: the traditional shop and the hard discount. For the other types of outlets, different situations arise which, however, are in line with what expected in relation to the nature of the asset (i.e. fresh products - fruit, vegetables and fish - are the most purchased in open markets). The exception is fresh meat for which traditional shop are more frequently visited as well as supermarkets and hypermarkets. Purchases 'via the internet' are actually sporadic excluding toys and, in part, coffee (especially in pods).

Figure 2.b.12 shows the types of outlet chosen by families in absolute poverty who have made at least one purchase in the two weeks of observation. Any comparison between these data and those reported



Products	No purchase - percentages of Households		
	No poor	Poor	Total
Bread	7.1	12.5	7.5
Pasta	18.8	32.1	19.7
Biscuits, rusks, snacks	17.7	31.2	18.6
Fresh meat	9.9	20.0	10.5
Frozen meat	25.6	39.9	26.5
Cured meats	16.6	41.8	18.2
Fresh fish	54.2	81.0	55.9
Frozen fish	63.8	80.6	64.9
Milk	19.6	31.1	20.3
Chesees	10.3	26.8	11.4
Yogurt	44.2	65.7	45.6
Eggs	27.8	38.8	28.5
Fres fruit	6.0	19.1	6.9
Fresh vegetables, potatoes and legumes	4.9	15.5	5.6
Dried or frozen vegetables, potatoes and legumes	44.0	63.3	45.3
Olive oil	58.5	75.9	59.6
Mineral water	29.0	48.1	30.3
Soft drinks	37.9	52.8	38.8
Wine	59.1	85.6	60.8
Coffee	42.7	65.2	44.2
Medicines	38.4	68.7	40.4
Personal hygiene products (soaps, deodorant, baby diapers, etc.)	25.1	48.7	26.6
Cleaning products	22.8	44.5	24.2
Disposable items for the kitchen (napkins, dishes, etc.)	41.1	63.1	42.5

Figure 2.b.10.: Families who did not purchase in the last two weeks (%) by product - Year 2019. Source: elaboration on 2019 Istat HES data.

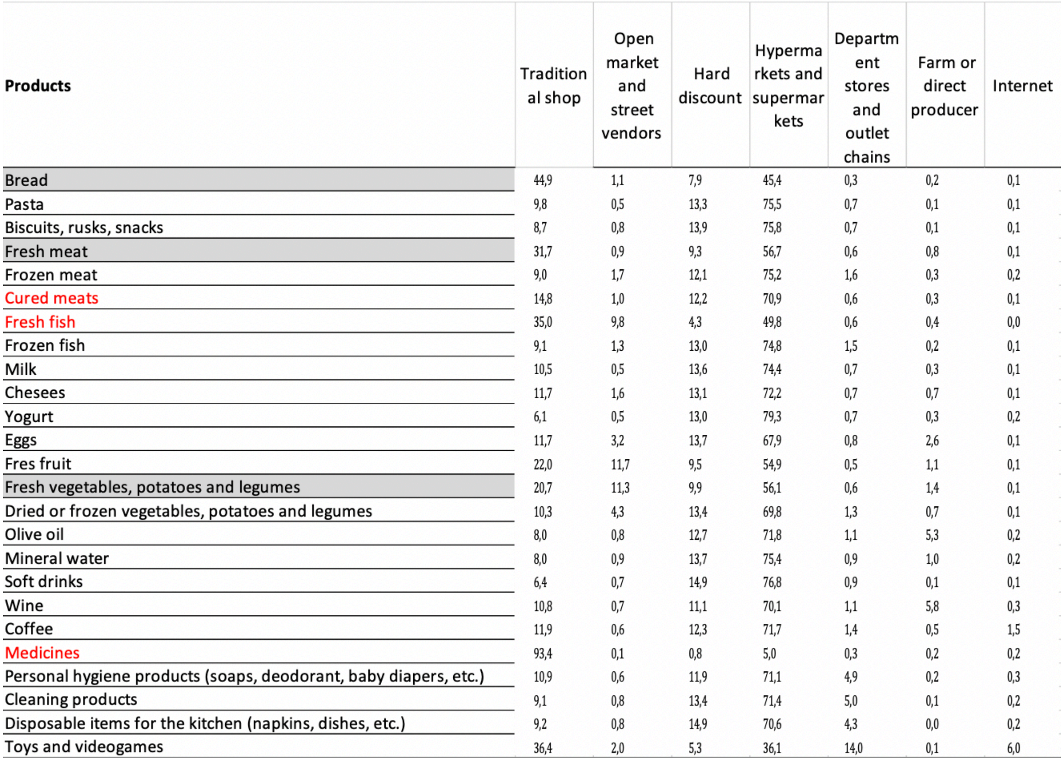
in figure 2.b.11 have to take into account the fact that the universe of poor households is small (6.4% in 2019) and that for many of the products in the table 'no purchase' is selected. Nevertheless, the main comments that it is possible to sketch are the following:

- On average, for the 25 products considered, only 10.6% of non-poor households made a purchase in a hard discount. This percentage reaches 27.2% (+ 16.6%) in the case of families in conditions of absolute poverty.

- The difference in the case of hypermarkets / supermarkets is almost identical but, of course, of the opposite sign (from 65.5% to 48.8%; - 16.7%).

- The difference is relevant for all products even if in the case of the products bought more frequently it tends to decrease. In the case of bread, for example, the use of the hard discount goes from 7.2% to 19.1% (+ 11.9%) and similarly occurs in the case of meat and vegetables.

- In general, the use of the traditional shop and open market is very similar for the two categories of families. On average, for the 25 products considered, 18.3% of non-poor households and 19.9% of poor ones make a purchase at a traditional store. For the open markets and street vendors the share is 2.3% and 2.9% respectively.



Products	Traditional shop	Open market and street vendors	Hard discount	Hypermarkets and supermarkets	Department stores and outlet chains	Farm or direct producer	Internet
Bread	44,9	1,1	7,9	45,4	0,3	0,2	0,1
Pasta	9,8	0,5	13,3	75,5	0,7	0,1	0,1
Biscuits, rusks, snacks	8,7	0,8	13,9	75,8	0,7	0,1	0,1
Fresh meat	31,7	0,9	9,3	56,7	0,6	0,8	0,1
Frozen meat	9,0	1,7	12,1	75,2	1,6	0,3	0,2
Cured meats	14,8	1,0	12,2	70,9	0,6	0,3	0,1
Fresh fish	35,0	9,8	4,3	49,8	0,6	0,4	0,0
Frozen fish	9,1	1,3	13,0	74,8	1,5	0,2	0,1
Milk	10,5	0,5	13,6	74,4	0,7	0,3	0,1
Chesees	11,7	1,6	13,1	72,2	0,7	0,7	0,1
Yogurt	6,1	0,5	13,0	79,3	0,7	0,3	0,2
Eggs	11,7	3,2	13,7	67,9	0,8	2,6	0,1
Fres fruit	22,0	11,7	9,5	54,9	0,5	1,1	0,1
Fresh vegetables, potatoes and legumes	20,7	11,3	9,9	56,1	0,6	1,4	0,1
Dried or frozen vegetables, potatoes and legumes	10,3	4,3	13,4	69,8	1,3	0,7	0,1
Olive oil	8,0	0,8	12,7	71,8	1,1	5,3	0,2
Mineral water	8,0	0,9	13,7	75,4	0,9	1,0	0,2
Soft drinks	6,4	0,7	14,9	76,8	0,9	0,1	0,1
Wine	10,8	0,7	11,1	70,1	1,1	5,8	0,3
Coffee	11,9	0,6	12,3	71,7	1,4	0,5	1,5
Medicines	93,4	0,1	0,8	5,0	0,3	0,2	0,2
Personal hygiene products (soaps, deodorant, baby diapers, etc.)	10,9	0,6	11,9	71,1	4,9	0,2	0,3
Cleaning products	9,1	0,8	13,4	71,4	5,0	0,1	0,2
Disposable items for the kitchen (napkins, dishes, etc.)	9,2	0,8	14,9	70,6	4,3	0,0	0,2
Toys and videogames	36,4	2,0	5,3	36,1	14,0	0,1	6,0

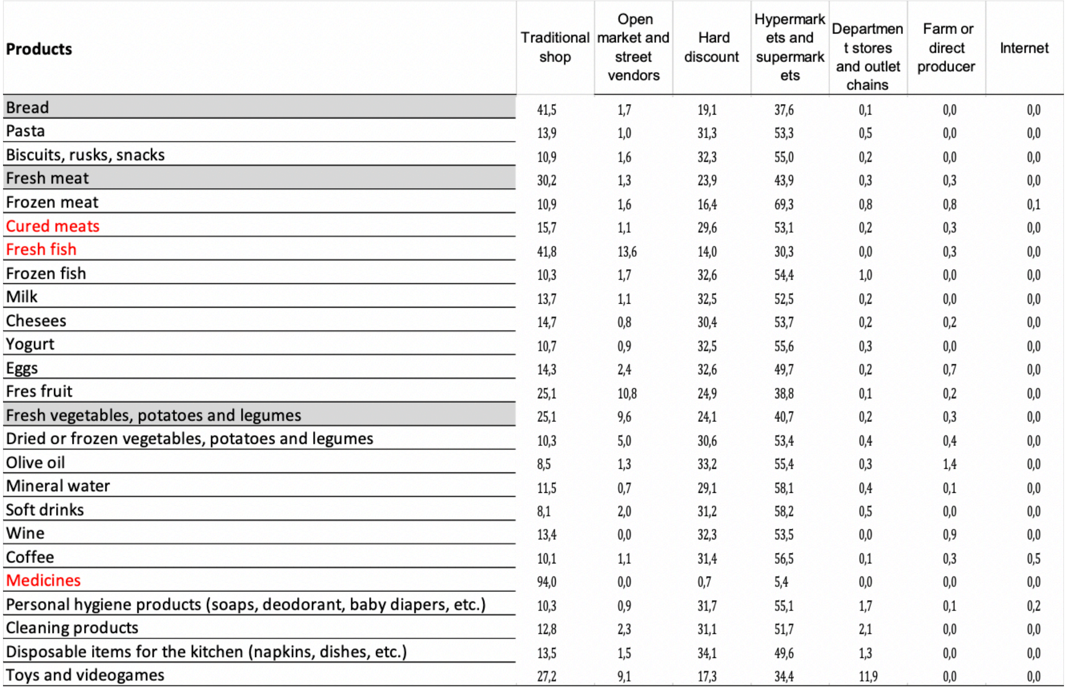
Figure 2.b.11.: Types of outlet where households make purchases (% distribution) - Year 2019. Source: elaboration on 2019 Istat HES data.

- The last three categories of businesses (department stores, farms and Internet) are almost completely unused by poor families (an expected result). If on average 3.3% of non-poor families buy from one of these types of outlets, in the case of the poor, the share drops to 1.2%.

The results obtained from these preliminary analyses of 2019 HES data show some interesting differences between non poor and absolutely poor households in terms of choice of the type of outlet where purchasing a list of 25 large consumption products. This evidence is worth to be deepened also by breaking down the analysis at territorial level, overcoming the problem of a too small sample if we take into consideration only poor households. This line of research is aimed at improving the estimation of the actual prices paid by the poor families in different Italian geographical areas by taking into account their different behavior in the choice of the outlet where purchasing in particular large consumption products. The possible results obtained could enhance the spatial comparison of consumer prices by making reference to the poor part of the population.

2.b.7. The impact of the local cost-of-living differences on the measure of the poverty incidence

Intra-country comparisons of poverty indicators are important for many reasons. For example, when measuring the poverty incidence, the use of a national poverty line allows to establish a general scheme of how local areas (e.g. regions or provinces) compare with national standards. However, considering the same poverty line for each area implies an equity concept in which individuals with equal income are assumed to have similar wellbeing regardless of the area where they live. The use of local poverty



Products	Traditional shop	Open market and street vendors	Hard discount	Hypermarkets and supermarkets	Department stores and outlet chains	Farm or direct producer	Internet
Bread	41,5	1,7	19,1	37,6	0,1	0,0	0,0
Pasta	13,9	1,0	31,3	53,3	0,5	0,0	0,0
Biscuits, rusks, snacks	10,9	1,6	32,3	55,0	0,2	0,0	0,0
Fresh meat	30,2	1,3	23,9	43,9	0,3	0,3	0,0
Frozen meat	10,9	1,6	16,4	69,3	0,8	0,8	0,1
Cured meats	15,7	1,1	29,6	53,1	0,2	0,3	0,0
Fresh fish	41,8	13,6	14,0	30,3	0,0	0,3	0,0
Frozen fish	10,3	1,7	32,6	54,4	1,0	0,0	0,0
Milk	13,7	1,1	32,5	52,5	0,2	0,0	0,0
Chesees	14,7	0,8	30,4	53,7	0,2	0,2	0,0
Yogurt	10,7	0,9	32,5	55,6	0,3	0,0	0,0
Eggs	14,3	2,4	32,6	49,7	0,2	0,7	0,0
Fres fruit	25,1	10,8	24,9	38,8	0,1	0,2	0,0
Fresh vegetables, potatoes and legumes	25,1	9,6	24,1	40,7	0,2	0,3	0,0
Dried or frozen vegetables, potatoes and legumes	10,3	5,0	30,6	53,4	0,4	0,4	0,0
Olive oil	8,5	1,3	33,2	55,4	0,3	1,4	0,0
Mineral water	11,5	0,7	29,1	58,1	0,4	0,1	0,0
Soft drinks	8,1	2,0	31,2	58,2	0,5	0,0	0,0
Wine	13,4	0,0	32,3	53,5	0,0	0,9	0,0
Coffee	10,1	1,1	31,4	56,5	0,1	0,3	0,5
Medicines	94,0	0,0	0,7	5,4	0,0	0,0	0,0
Personal hygiene products (soaps, deodorant, baby diapers, etc.)	10,3	0,9	31,7	55,1	1,7	0,1	0,2
Cleaning products	12,8	2,3	31,1	51,7	2,1	0,0	0,0
Disposable items for the kitchen (napkins, dishes, etc.)	13,5	1,5	34,1	49,6	1,3	0,0	0,0
Toys and videogames	27,2	9,1	17,3	34,4	11,9	0,0	0,0

Figure 2.b.12.: Types of outlet where poor households make purchases (% distribution) - Year 2019. Source: elaboration on 2019 Istat HES data.

lines allows to gauge intra-country poverty, which can be important for planning local policies.

A possibile approach to compute local poverty lines is by taking into account the different price levels within the country. Under this approach the national poverty line can be modified using area-specific Purchasing Power Parities (PPPs), following the methodology currently applied at international level for international comparisons among different countries (see section 2.b.3). In this section we show an application where we compute the Head Count Ratio – a measure of poverty incidence – using Household Expenditure Survey data in Italy, adjusting the national poverty line using the SPI($Q_{0.2}$) values computed using the methodology presented in section 2.b.4.

In this application we use HES data as they are internationally used – together with EU-SILC data – to compute monetary poverty indicators, such as the HCR. It is important to underline that, therefore, the current application could be extended to other countries and/or datasets, as it presents a general methodology to compute local poverty lines.

According to ISTAT, the Head Count Ratio (HCR), a relative measure of poverty incidence, is computed using HES consumption data by defining for each household an indicator variable which takes value 1 if the Monthly Consumption Expenditure (MCE) of the household is less or equal the poverty line, value 0 otherwise. The values are then averaged by using the sample weights. To compute the HCR values, it is thus necessary to first compute the poverty line. At national level, the poverty line for households of two components is set equal to the per-capita mean MCE at country level:

$$nPL = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_j} CE_{ij}w_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_j} a_{ij}w_{ij}} \tag{2.b.10}$$

where CE_{ij} represent the Consumption Expenditure, w_{ij} the survey weight and a_{ij} the household size of household j living in area i, with $i = 1, \dots, m$ and $j = 1, \dots, n_j$. To take into account the existence of economies of scale in consumption within households, the poverty line is then adjusted by using the Carbonaro scale (Istat, 2010). In this way, household expenditures can be directly compared with those of households composed of two members. The value of the HCR_{ij} is thus computed for each household as

$$HCR_{ij} = I(CE_{ij} \leq PL \cdot s_{ij}) \tag{2.b.11}$$

where s_ij represents the values of the Carbonaro scale, a specific coefficient depending on the household size. Specifically, according to the Carbonaro scale $s_ij = 0.66$ for households with $a_{ij=1}$, $s_{ij} = 1.33$ for a household with $a_{ij} = 3$, $s_{ij} = 1.63$ when $a_{ij} = 4$, $s_{ij} = 1.90$ when $a_{ij} = 5$, $s_{ij} = 2.16$ when $a_{ij} = 6$ and $s_{ij} = 2.40$ for households with 7 members or more. The *HCR* of a given area i computed by using the national poverty line PL is then computed as

$$HCR_i = \frac{\sum_{j=1}^{n_j} HCR_{ij}w_{ij}}{\sum_{j=1}^{n_j} w_{ij}}. \tag{2.b.12}$$

A corresponding measure of variability can be computed to derive the coefficient of variation and the confidence intervals for the *HCR* estimates. We computed direct estimates using the *sae* package that is available in *R* (Molina and Marhuenda, 2015).

To allow intra-country comparisons, local poverty lines can be computed and used in the HCR definition. A possibility to compute local poverty lines is by taking into account the different price levels in each area, modifying the national poverty line using area-specific Purchasing Power Parities (PPPs), so that the poverty lines represent approximately the same standard of living across the different areas. By considering the provincial SPI(s values computed using the ASESD methodology and using mean prices (see section 2.b.4), the national poverty line can be adjusted for each province using the SPI($Q_{0.2}$) values opportunely weighted (adapting the idea in Renwick et al. (2014)):

$$nPL_i^* = nPL \times (\lambda_i SPI_i + 1 - \lambda_i) \tag{2.b.13}$$

where nPL_i^* is the adjusted poverty line for province i, λ_i is the estimated share of food consumption in province i. The quantities λ_i's are estimated from the HES 2017 as the provincial mean of the ratios between the rent expenditure and the total consumption expenditure:

$$\lambda_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} \frac{p_{ij}}{t_{ij}} w_{ij}, \tag{2.b.14}$$

where n_i is the sample size in province i, w_{ij} is the survey weight of household j in area i, p_{ij} is the food expenditure of household j in area i and t_{ij} is the total consumption expenditure of household j

in area i. The survey weights have been calibrated to sum to the total households at provincial level.

Although the λ_i's are estimated at the provincial level – thus possibly unreliable because of small sample size – we judge the direct estimates suitable for our purpose. Indeed, about half of the provinces have a 95% confidence interval for λ_i%'s direct estimates that is less than 4% and it is less than 5% for about 75% of the provinces[4]. In table 2.b.6 we show the distribution over provinces of the λ_i's grouped by the main Italian geographic areas, which is similar among provinces in the north, center and south of Italy.

Table 2.b.6.: Distribution over provinces of the λ_i%'s grouped by Italian geographic repartitions

Repartition	Min	1st Q.	Median	Mean	3rd Q.	Max
North	12.89	16.36	17.74	18.04	18.94	26.83
Centre	13.87	18.05	19.11	19.38	20.29	25.47
South	18.96	21.34	23.71	23.34	25.27	27.74

From Table 2.b.6 we can see that distribution of the share of expenditure for food is higher in the southern provinces than in the central and northern provinces, with a mean value equal approximately to 20%.

Having computed the adjusted nPLs, we then calculated the corresponding direct estimates of the poverty rates. We computed the direct estimates using the `direct` function of the R (R Core Team, 2019b) package `sae` (Molina and Marhuenda, 2015). We judged the variability of the direct estimates to be too high, in particular to carry out comparisons among provinces. Indeed, about half of the provinces had a 95% confidence interval length grater than 6%, and about one third of them greater than 9%. Looking at the coefficient of variation (CV) of the direct estimates, we obtained for approximately half of the provinces a CV greater than 30% and for about 25% of the provinces a CV greater than 45%. Therefore, we decided to resort to small area estimation methodologies to try to improve the efficiency of the poverty incidence estimates.

We used a Fay-Herriot (FH) model for the HCR at provincial level in Italy using the adjusted poverty lines (used to compute direct estimates). As auxiliary variables we used the ratio between number of taxed persons over the population, and the ratios between the number of persons with *i.* income coming from salary, *ii.* income coming from pensions and *iii.* income lower than 10,000 euros per year, over the number of taxed persons. These data come from the Italian tax agency database 2017.

The EBLUPs (Empirical Best Linear Unbiased Predictors) obtained with the FH model showed a gain in efficiency with respect to direct estimates. We obtained a CV smaller than 16% in 37 provinces, while half of the provinces had a CV smaller than 20%. The gain in term of variability is shown in figure 2.b.13 where we can see that the EBLUP is more efficient than the Direct estimator in all the provinces and the gain in efficiency is greater in those areas where the sample size is smaller than expected.

[4] Standard error of λ_i's are obtained ignoring the design effect at the province level.

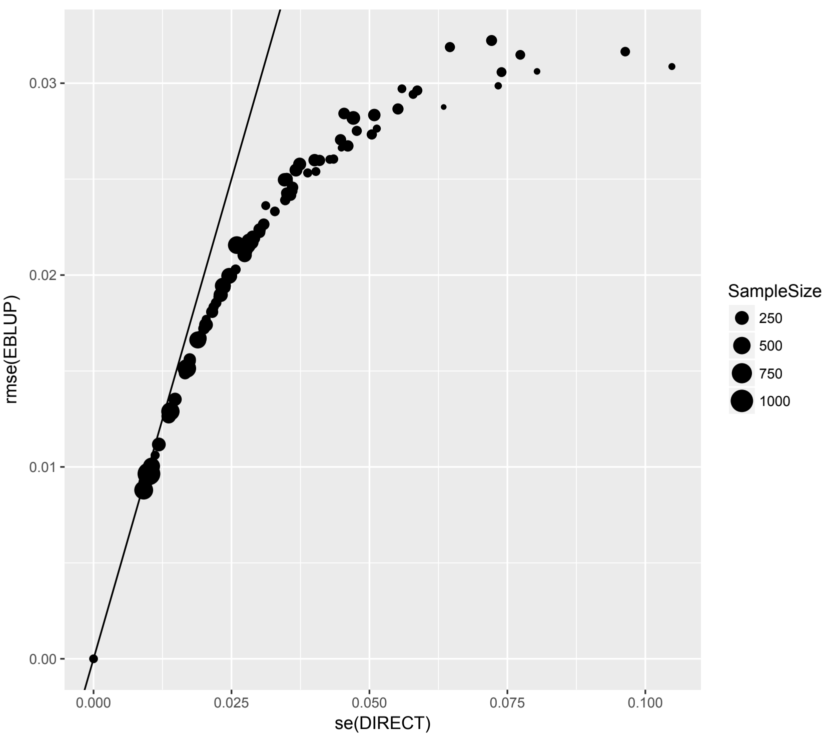


Figure 2.b.13.: Comparison of the Root Mean Squared Error of EBLUP and direct estimates.

Figure 2.b.14 maps the HCR computed using the price-adjusted poverty lines referring to the Italian provinces[5]. As we can see, the results confirm the well-known north/south divide, with HCR values that are generally higher in the south of the country, lower in the north.

[5] The HCR value has not been estimated for out of sample provinces in the HES data. Specific predictions could be made for this provinces.

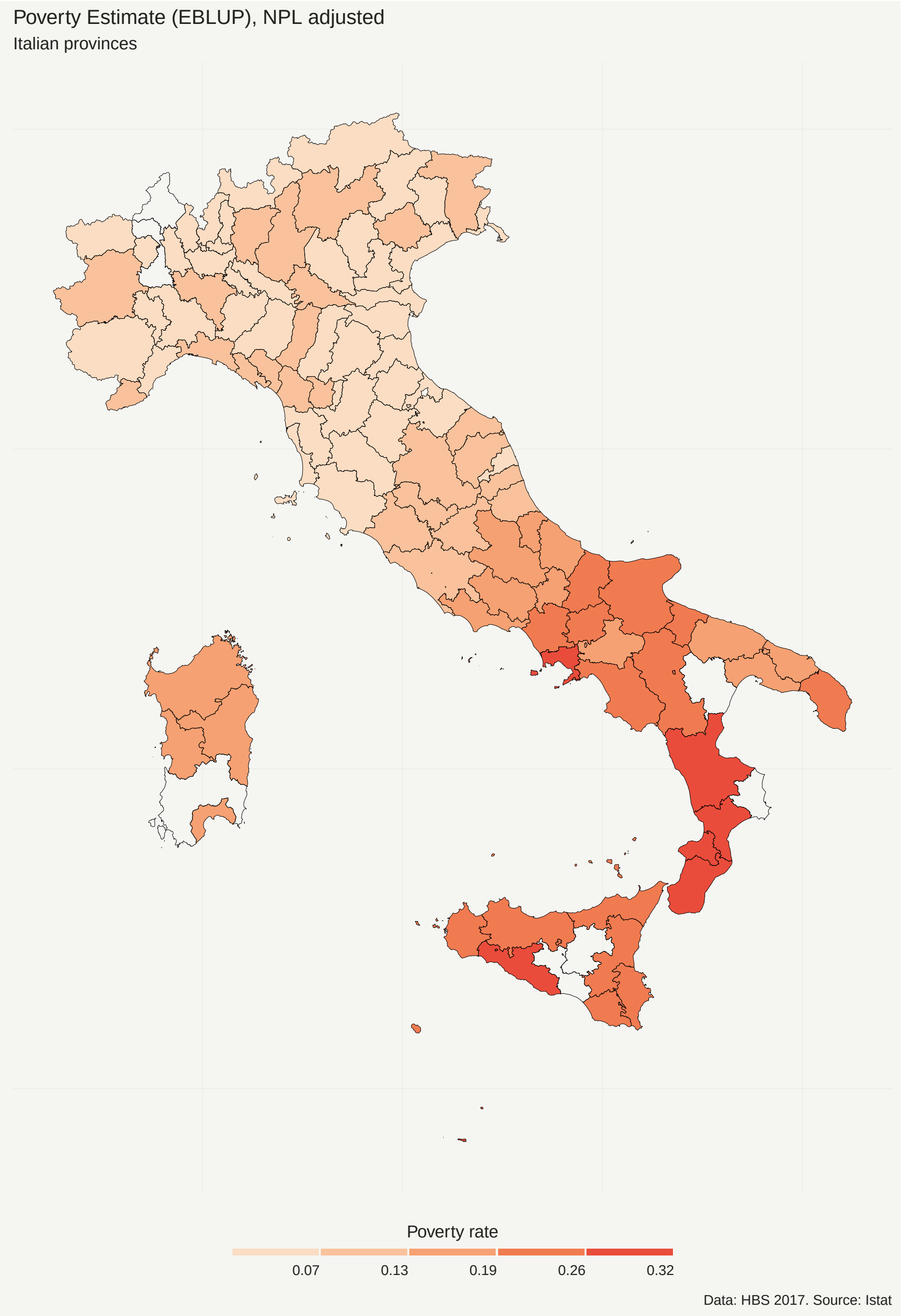

Poverty Estimate (EBLUP), NPL adjusted
Italian provinces
Poverty rate
0.07
0.13
0.19
0.26
0.32
Data: HBS 2017. Source: Istat

Figure 2.b.14.: Poverty rate at provincial level in Italy: EBLUPs computed using the adjusted national poverty line.

We also computed the EBLUPs without any adjustment of the national poverty line, using the same small area model as for adjusted EBLUPs. In this case, however, the map of the HCR using the national poverty line is essentially the same as the one using the price-adjusted poverty lines, as the SPI($Q_{0.2}$) are applied only to approximately the 20% of the poverty line, as explained above. A finer comparison is represented in Figure 2.b.15, where we can see that using the SPI($Q_{0.2}$) to adjust the poverty lines, the HCRs in northern and central provinces slightly decrease.

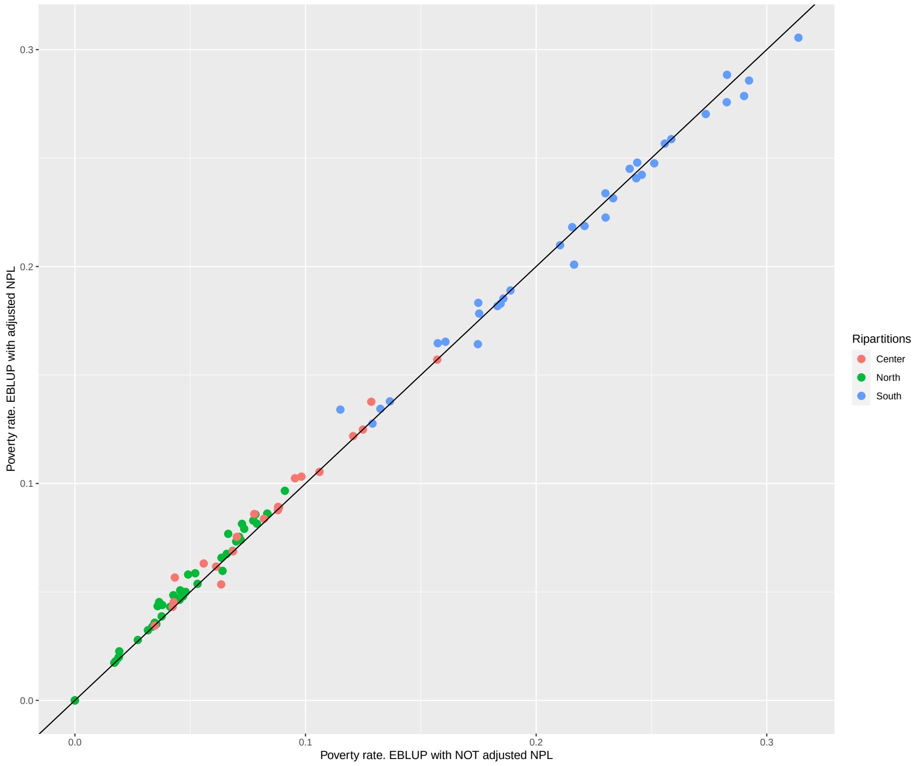


Figure 2.b.15.: Poverty rate at provincial level in Italy: provincial EBLUPs estimates using the SPI($Q_{0.2}$) adjusted vs not adjusted national poverty line.

The results obtained here suggest that the methodology can be extended to include other Spatial Price Indexes, therefore adjusting the national poverty line with other components of households' consumption expenditure. Indeed, our results suggest the products included in the scanner data represent a relevant but still limited share of the total household consumption expenditure, approximately equal to 20%. Therefore, by including other consumption expenditure components, such as for example the expenditure for rent, the national poverty line could be adjusted in a more complete manner. Figure 2.b.16 reports the same analysis as figure 2.b.15 but also using, in addition to the SPI($Q_{0.2}$), a Spatial Rent Index (SRI) to further adjust the national poverty line. The SRI has been estimated using HES data and small area estimation methodologies. It is important to underline that the cost for the rent covers, in mean, another 20% of the total household consumption expenditure. From figure 2.b.16 we can see that the effect of the adjustment in this case is more pronounced.

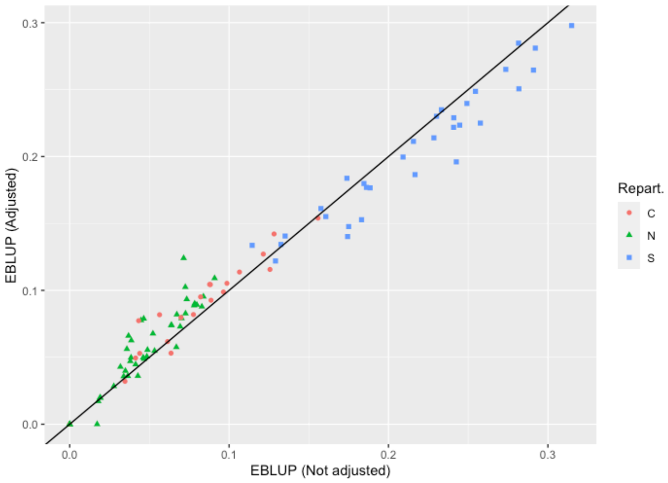


Figure 2.b.16.: Poverty rate at provincial level in Italy: provincial EBLUPs estimates using the national poverty line adjusted with the SPI($Q_{0.2}$) and SRIs vs EBLUPs estimates using the not adjusted national poverty line.

Finally, as already stressed, it is important to underline that, being based on EU–SILC and HES data, the proposed analyses - applied here only to Italian data - could be extended to other European countries provided that survey data at subnational level are available.

2.b.8. Final remarks

Integration of traditional data with Big data sources is the red line followed by many modern statistical agencies in the production of official data. Here we have described the assumptions and the steps necessary to innovate the production of poverty indicators, which are relevant SDGs indicators, using scanner data on prices of RTCs.

We are convinced that two major innovations are important when focusing on the possible estimates of people vulnerability: the first one concerns the inclusion of the measurement of cost of living or regional inflaction in these, the second is extending their geographical notation to offer measures related to the places where people live. This means allowing for estimates which refer also at a subregional level individuated by NUTS3 level in European classifications.

The results of our study are useful to either looking at the measurement of poverty and inequality and/or to the measurement of regional inflation.

The proposed methodology is applicable in European countries as it is based on current sample surveys as EU–SILC and HES and on scanner data on prices of RTCs, that are generally available for NSIs in western countries. The approach is model–based and uses tools which are now in the current tool–box of the majority of NSIs in European Statistical System for the analysis of price data (the CPD models) and of survey data (SAE models).

Integrating data sources is generally a complex process, for this it can be useful to list recommendations which stem from our analysis:

- price distributions for groups of items (BHs or other groupings) obtained by scanner data are a valuable information. The first quantiles (quintiles, or deciles) of these distributions are mimicking the prices paid by the poor. This is the evidence from Section 2.b.6 where we see that with reference to the poor, a suitable place for shopping is the hard discount.

- the SPIs (f.i. SPI($Q_{0.2}$)) alone are not enough to appreciate the subregional variations of the cost of living, they should be accompanied by the SPIs for housing Rents (SPI Rs) obtained by HES or by archives of Revenue and Tax Agencies.

- the prices from scanner data may be affected by the price policy of the RTCs, this can have a uniform effect at the country level, smoothing the subregional variability of the prices. In this, distinguishing between the so–called first–price products and the others can facilitate the analysis. The first–price products group all the products with the lowest price that exists for each product category in a supermarket assortment.

- integration of different data sources as done in this study requires for a deep reflection on the sources of uncertainty affecting the whole process. They come from the design of the data sources and from the application of models as CPD and SAE models. Scanner data bases can be the results of probability or not probability sampling of items or observations (as in the design of the Istat data base of prices from scanner data). While the accuracy of the model-based estimates has already been studied in the literature, the effect of the design of the data sources, especially in case of Big data sources, has not yet been completely studied. This last topic in combination with the accuracy of model-based estimates necessitates further research.

2.c. Estimation of local wealth using remote sensing data

In Deliverable 2.2 we reviewed literature on the use of new forms of data and alternative spatial data sources for producing small area estimates of SDG related indicators. The terms new forms of data and alternative sources of data will be used interchangeably. In this Section we use remotely sensed data in an application of small area estimation for producing estimates of average wealth at the level of Upazillas in Bangladesh. Although mobile-phone data are available in this case, the data licence does not allow their use in this application.

As mentioned in Deliverable 2.2 emphasis is placed upon area-level models since the literature that utilises new forms data does so by aggregating the data at some level of spatial scale. The use of area-level models in conjunction with news forms of data provides a powerful combination that reduces the need to rely on Census and unit-level data. Areal-level models offer a feasible approach to the production of small area statistics. However, the specification of area-level models requires technical knowledge in order to avoid commonly occurring mistakes. The present deliverable has two aims (a) to illustrate the use of remote sensing data as auxiliary information in small area estimation and (b) to illustrate and discuss potential pitfalls resulting from the use of automated algorithms. At this point we must clarify that these pitfalls are due to the possible lack of detailed understanding of the models not because of the algorithms, which are powerful when used correctly.

This section is organised as follows. In the next subsection we review area-level models for small area estimation. Subsection 2.c.2 presents a small area application using remote sensing data as auxiliary information. Subsection 2.c.3 presents a detailed analysis of potential pitfalls when specifying small area models using open source software. The last subsection summarises the main findings and outlines areas for future work.

2.c.1. Area-level models

In this subsection we review small area estimation (SAE) methods before focusing on the application of interest. In many applications direct estimation, that is estimation using only area-specific data, leads to unreliable estimates due to the small domain-specific sample sizes and the associated high sampling variability. In such cases one has to rely upon alternative model-based methods for producing small area estimates. Model-based methods assume a model for the population and sample data and construct optimal predictors of the target parameters under the model. The term predictor instead of estimator is conventionally used as, under the model, the target parameters are assumed to be random. Generally speaking, SAE models can be classified in two broad classes, namely unit-level models and area-level models(Fay and Herriot, 1979). Access to data for fitting area-level models is easier because of less strict confidentiality constraints. Moreover, many of the applications that explore the use of new forms of data use area-level models because satellite and mobile phone data are available in aggregate form at an acceptable geographical level.

The starting point for area level models is an unbiased direct estimator of the target parameter. Denoting by y_{ij} the outcome of interest for a unit j in area i and by n_i denote the sample size for area

i, the simplest direct estimator of the population average or proportions the Hajek-Brewer estimator which is defined as follows,

$$\hat{\theta_i}^{Direct} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

with w_{ij} denoting the survey weights. Provided that we have access to information about the design of the survey, the variance of the direct estimate can be computed by using standard survey estimation techniques. For example, the variance of the direct estimates can be estimated using analytic and replication methods depending on the design of the survey and the information available to the data analyst. In the application we present in this deliverable we use an ultimate cluster variance estimator (Heering et al., 2017), Section 3.6.1 as an approximation to a multistage design. As we describe below, point and variance estimates of direct estimates are key parts for defining an area-level model.

Assuming that the variance of the direct estimate (sampling variance) is known, the area-level model -also known as the Fay-Herriot (FH) model can now be defined, and it is based on two stages. The first stage models the sampling variation with the sampling errors ϵ_i assumed to be independent and normally distributed $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$.

$$\hat{\theta_i}^{Direct} = \theta_i + \epsilon_i.$$

The second stage of the model assumes a linear model for θ_i,

$$\theta_i = x_i^T \boldsymbol{\beta} + u_i,$$

where x_i^T denotes the $i-th$ area covariates, $\boldsymbol{\beta}$ denotes the regression parameter vector and u_i represents the area-specific random effects which are assumed to be also normally distributed, $u_i \sim N(0, \sigma^2_u)$. Combining the two models defines the FH model,

$$\hat{\theta_i}^{Direct} = x_i^T \boldsymbol{\beta} + u_i + \epsilon_i.$$

At this point it is important to understand that with the area level model we need to assume that $\sigma^2_{\epsilon_i}$ is known otherwise it would not be possible to identify both variance parameters because both terms reflect variation at the same level. Correctly allocating the variability to the two levels has an impact on the prediction of the random effects and hence on point and mean squared error estimation. Finally, assuming that the sampling variances are area-specific is a reasonable assumption given that the sample sizes may vary widely between areas. Although these are technical details, they are important ones for the correct specification of area-level models.

Estimates of $\boldsymbol{\beta}$, u_i and σ^2_u are then obtained by using maximum likelihood, residual maximum likelihood or Bayesian methods that are available via *R* packages such as the *sae*, *emdi* and *brugs* packages. Using the estimates of the regression parameters, the estimated variance component and the predicted random effect, small area predictors of the target population parameter can then be derived using the following expression

$$\hat{\theta}_i^{FH} = x_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i = \hat{\gamma}_i \hat{\theta}_i^{direct} + (1 - \hat{\gamma}_i) x_i^T \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{(\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon_i}^2)}$.

As we discussed in Deliverable 2, a number of papers that use so-called alternative forms of data propose methods similar in spirit to mainstream SAE literature (e.g. Steele et al., 2017). In particular, Steele et al. (2017) present methodology for poverty mapping in Bangladesh that uses mobile and satellite auxiliary data when Census data are out-of-date or unavailable. The paper uses area-level models with auxiliary variables derived from call detail records (CDRs) and remote sensing (RS) data. CDR data are used for extracting metrics such as phone usage, top up amounts and network information related to mobile phone usage. RS data include metrics likely to be associated with wealth indicators and include night-time lights, vegetation indices, climatic conditions and distance from roads and major urban areas. In the application we present in this deliverable we employ part of the data used by Steele et al. (2017). In particular, we produce estimates of an average wealth index for Upazillas in Bangladesh by modelling this index as a function of RS data. For the current application we use a hierarchical Bayes (HB) framework because much of the applied work with big data uses Bayesian-type estimation tools for example approximate Bayesian methods such as Integrated Nested Laplace Approximation (INLA). This choice is justified by the fact that Bayesian methods offer a more straightforward approach than frequentist methods to evaluating the uncertainty of the small area estimates via the posterior distribution and complex models for example those involving spatially correlated random effects can be estimated in a more straightforward way. From a hierarchical Bayesian perspective the FH model is defined as follows,

$$\hat{\theta}_i^{Direct} | \theta_i \sim N(\theta_i, \sigma_{\epsilon_i}^2)$$
$$\theta_i | \beta, \sigma_u^2 \sim N(x_i^T \beta, \sigma_u^2)$$

The prior distributions for β and σ_u^2 are usually diffuse, for instance, an improper uniform distribution or a normal distribution with large variance for β and an inverse gamma for the precision parameter $1/\sigma_u^2$. Careful use of Bayesian methods via automated algorithmic tools is of paramount importance. For example, when working with area-level models, the analyst must be careful as to how he/she handles the sampling variances of the direct estimates, $\sigma_{\epsilon_i}^2$. As discussed in the previous section, $\sigma_{\epsilon_i}^2$ is either estimated using survey micro-data or is given by the data provider and -in the simplest case- is assumed to be known (fixed). Although for the purposes of producing small area estimates of the wealth index for Upazilas we will treat $\sigma_{\epsilon_i}^2$, one can capture this additional uncertainty due to the estimation of the sampling variances by including another hierarchical level (see You and Chapman (2006)) in the hierarchical Bayes model. In applied work, outside the mainstream small area literature, it is not clear how the sampling variances are treated when automated (black box -type) algorithms are used. The issue of how the sampling variances are estimated and treated also relates to how the analyst decides to specify the target geography. In addition to presenting an application using remotely sensed data, in this deliverable we discuss such issues, assess the impact on small area estimates and propose possible solutions.

Estimates of the uncertainty of the small area estimates are also required for assessing the quality of the small area estimates. Uncertainty in this case is quantified by the estimated Mean Squared Error that can be obtained both analytically using for example a Prasad-Rao estimator (Prasad and Rao, 1990) or by using parametric bootstrap under the area-level model. The *R* packages *sae* and *emdi* include both analytic and parametric bootstrap MSE. For estimates produced using the HB framework or INLA MSE estimates are produced by using the posterior distribution of the small area estimates.

2.c.2. Application

The application we present as part of this deliverable comprises the production of estimates of wealth- quantified by the average of a wealth index (WI)- for Upazilas (administrative level 3) in Bangladesh. The survey data we use come from the 2014 Demographic and Health Survey (DHS). The DHS in Bangladesh uses a stratified 2-stage cluster design with at least one cluster selected in 365/508 (72%) Upazilas. In total there are 17,000 households with the average number of households at the level of Upazila equal to 34. The DHS WI is constructed by taking the first principal component of a basket of household assets and housing characteristics such as floor type and ceiling material, which explains the largest percentage of the total variance, adjusting for differences in urban and rural strata. A final composite combined score is then used as a WI where each household is assigned its correspondent quintile in the distribution and each individual belonging to the same household shares the same WI score. A higher score implies higher socioeconomic status. Here, we use aggregated average WI scores per Upazila as the direct estimate in the FH model. The covariates we use come from remote sensing and include information about an enhanced vegetation index, night time lights , elevation and accessibility to areas with more than 50K people all aggregated at Upazila level. Additional details about the dataset we use can be found in Steele et al. (2017).

2.c.2.1. Direct and model-based small area estimates of wealth using remote sensing data

The first step is to produce direct and model-based estimates, using the Fay-Herriot model, of the mean wealth index in each upazila and associated variance and mean squared error estimates. The direct estimates of WI at the level of Upazila are produced using the survey weighted estimator of the average WI, with the DHS weights, and are presented in Figure 2.c.1. The variance of the direct estimates is produced by using the ultimate cluster variance (UCV) approach which offers an approximation to the survey design of the DHS in Bangladesh. For those upazilas comprising one cluster, we used a design factor (DEFT) for producing variance estimates since the ultimate cluster variance approach is not applicable in those cases. The variances of the direct estimates are presented on the left plot in Figure 2.c.2.

Model-based estimates using the Fay-Herriot model are produced using four remote sensing covariates described in the introduction of the application section. The Fay-Herriot model assumes the knowledge of the variances of the direct estimates which in the simplest case are assumed to be fixed. The variances of the direct estimates are estimated by using the UCV approach and additionally smoothed using the generalised variance function (GVF) approach. In our case, the use of the GVF approach assumes a model for the design variances of the estimated average wealth index in each Upazila as a function of polynomial terms of the Upazila-specific sample sizes. Fay-Herriot model estimates are produced with the R package *sae* and plotted against the direct estimates in Figure 2.c.1. The results show that the model-based estimates using only remote sensing covariates are well correlated with direct estimates. The advantage of the model we consider in this deliverable is that it does not rely on the availability of Census data. Estimates of the mean squared error of the Fay-Herriot point estimates are produced by using the Prasad-Rao estimator. MSE estimates are plotted against Figure 2.c.2. The results show that on average model-based estimates have lower MSE than the variance of the direct estimates.

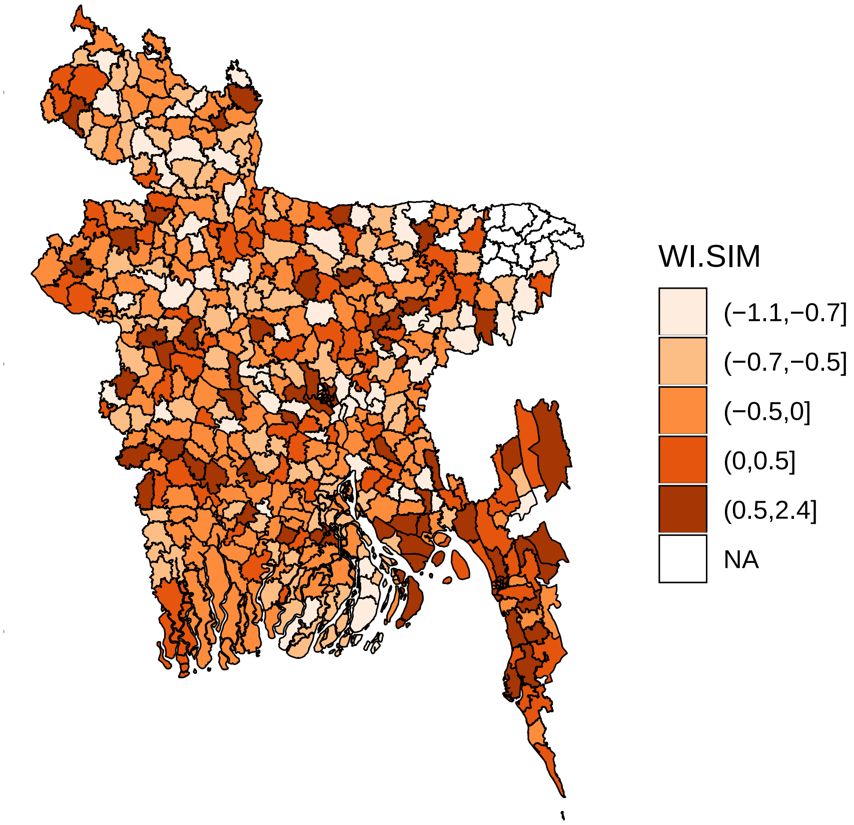


Figure 2.c.1.: Direct estimates of WI for Upazilas in Bangladesh using the DHS 2014 data

2.c.3. Pitfalls in model-based estimation

In the previous subsection we illustrated the production of model-based estimates using auxiliary information from remote sensing data. In this section we illustrate some of the pitfalls in using model-based methods. Research work outside the mainstream small area literature that uses remote sensing data as auxiliary information relies on the uses of approximate Bayesian inference to estimates aggregate-level models. More specifically, models are estimated using Integrated Nested Laplace Approximations (INLA). Inference for the model-based estimates is done using the posterior distribution of the model estimates. It is surprising to see that research papers on the use of big-data are not discussing in detail the specification of the model. Nevertheless, the specification of aggregate (area)-level models involves important technical details such as the estimation and treatment of the variances of the direct estimates. In Steele et al. (2017) models were built on the scale of Voronoi polygons (see Section 2.1 in

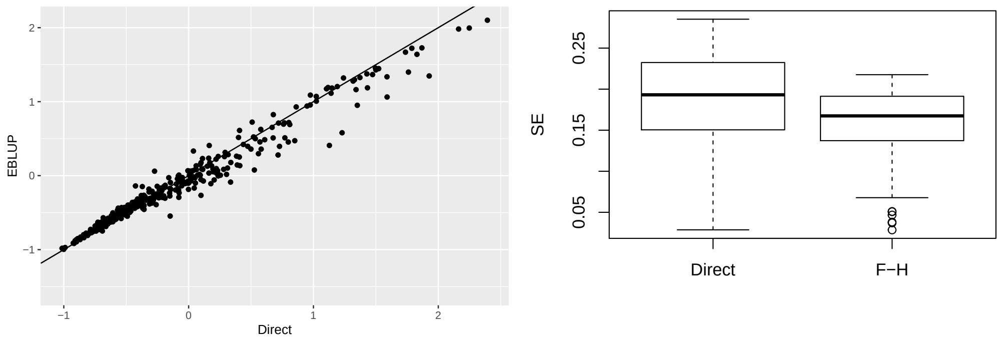


Figure 2.c.2.: Direct and model-based (FH) point and variance of the direct and MSE of the model-based estimates of the mean wealth index for Upazilas

the corresponding paper). The choice is justified by the inclusion of covariates obtained from mobile phone data that are associated with a geography based on the location of the communication towers. However, it is not clear how the model is specified and how the variances of the direct estimates of the wealth index are treated. This is concerning because the use of INLA for fitting the models will converge to a solution even if the variance of the direct estimates is assumed to be unknown. We investigate this in more detail using the scenarios below.

Under scenario M1 we produce small area estimates under a frequentist perspective using the *R* package *sae*. The scenarios are summarised below. When using INLA in *R* to estimate the FH model, it is not clear how $\sigma^2_{\epsilon_i}$ is specified so that this is treated as known and fixed. For the purposes of our investigation we use *R* code from Steele et al. (2017) which shows sets the latent specification in INLA to be Gaussian i.i.d will set $\sigma^2_{\epsilon_i} = \sigma^2_{\epsilon}$. We call this the M2 scenario below. Clearly this is not appropriate as it does not account for the fact that areas have varying sample sizes. To the best of our knowledge, this issue is not discussed in applied work published outside the mainstream small area literature. Two other scenarios are tested. The first is trying to fix the issue within INLA by scaling the common level 1 variance by the ratio of area specific direct variances (M3). The second scenario (M4) specifies a hierarchical Bayesian model and fits this by using *BRUGS* in *R*. Notice that the prior distribution chosen for the precision parameters under models M3 and M4 is informative. The scenarios are summarised below.

M1 Standard FH model using `sae`. $\sigma^2_i = \hat{\sigma}^2_i$ fixed

M2 Standard Gaussian model in R-INLA. $\sigma^2_i = \sigma^2_e$ unknown

M3 R-INLA with $\sigma^2_{e_i} = g_i \sigma^2_e$; $g_i = v_i / \bar{v}_i$.
$\tau = 1/\sigma^2_e$; $\pi(\tau) \sim \text{Gamma}\Big(\frac{\bar{n}_i - 1}{2} - 1\,,\ \frac{(\bar{n}_i-1)\bar{v}_i}{2}\Big)$.

M4 HB using `BRugs` with $\pi(\tau_i)$ as in M3.

Estimates of the fixed effects and the variance components under the four model specifications (M1-M4) are presented in Table 2.c.1. We observe that under all model specifications the estimates of the fixed effects are very close. However, large differences are observed in the estimates of the variance components between specification M1 (standard FH model) and M2 (the naive INLA specification) that fails to properly specify the level 1 variance. In particular, under the naive INLA specification (M2) estimated variance components are markedly different to the estimates obtained under M1 and the between area variance components is higher than the level 1 variance. This indicates that under M2 there is an issue with estimating the variance components. In contrast, under the adjusted INLA specification (M3) and the HB specification (M4) estimation of the variance components is improved and is close to the variance components estimated under the FH model (M1).

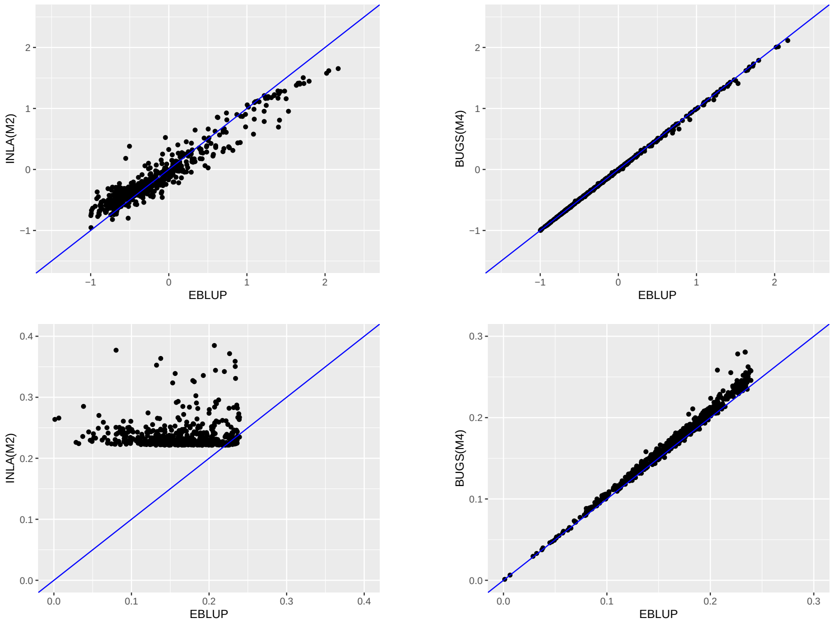
Table 2.c.1.: Estimates of the fixed effects and variance components under the four model specifications M1-M4.

	M1	M2	M3	M4
$\hat{\beta}_0$	0.6662	0.6866	0.6604	0.6469
$\hat{\beta}_{elev}$	-0.0553	-0.0530	-0.0557	-0.0548
$\hat{\beta}_{nl}$	0.3137	0.3112	0.3141	0.3159
$\hat{\beta}_{acc}$	-0.0878	-0.0892	-0.0874	-0.0838
$\hat{\sigma}_e^2$	0.0362	0.1219	0.0423	0.0408
$\hat{\sigma}_u^2$	0.1889	0.1058	0.1800	0.1838

A better insight to these results can be obtained by examining the point and MSE estimates under the different model specifications. Comparisons between the different sets of point estimates are presented in the top row of Figure 2.c.3. Generally, EBLUP point estimates are well correlated with point estimates under specifications M2,M3 and M4 but it is clear that the correlation is stronger for specifications M3 and M4. In papers that use big data sources for small area estimation it is customary to present scatterplots that correlate the point estimates produced with the big data sources and algorithms such as INLA and specification M1 to point estimates produced with Census data and conventional fitting algorithms. As illustrated above, estimates of the fixed effects under the different model specifications are quite close hence it is not surprising the area-specific estimates are well correlated despite the fact that the estimated variance components under M1 are different. Here we argue that such comparisons, especially those based on the use of area-level models, can be misleading. This can be seen by examining the second row of plots in Figure 2.c.3. The plots show the correlation between the square root of the MSE estimates of the EBLUP obtained using the Prasad-Rao estimator and the square root of the MSE estimates under specifications M2,M3 and M4 using the posterior distribution. The results show that the MSE estimates under M2 bear no relation to those of the EBLUP. This is to be expected due to the problems with estimating the variance components under M2. The correlation improves under M3 and is clearly better under the properly specified HB model M4. The last set of results demonstrates that looking just at correlations of point estimates can be misleading and can mask problems with the model specification. Comparing the MSE estimates can help with uncovering issues in the specification of a model.

Further insights into the problems with model specification M2 are offered by the following sensitivity

Figure 2.c.3.: Comparisons of point (first row) and MSE (second row) estimates under the FH model (EBLUP) and the M2,M3 and M4 model specifications.

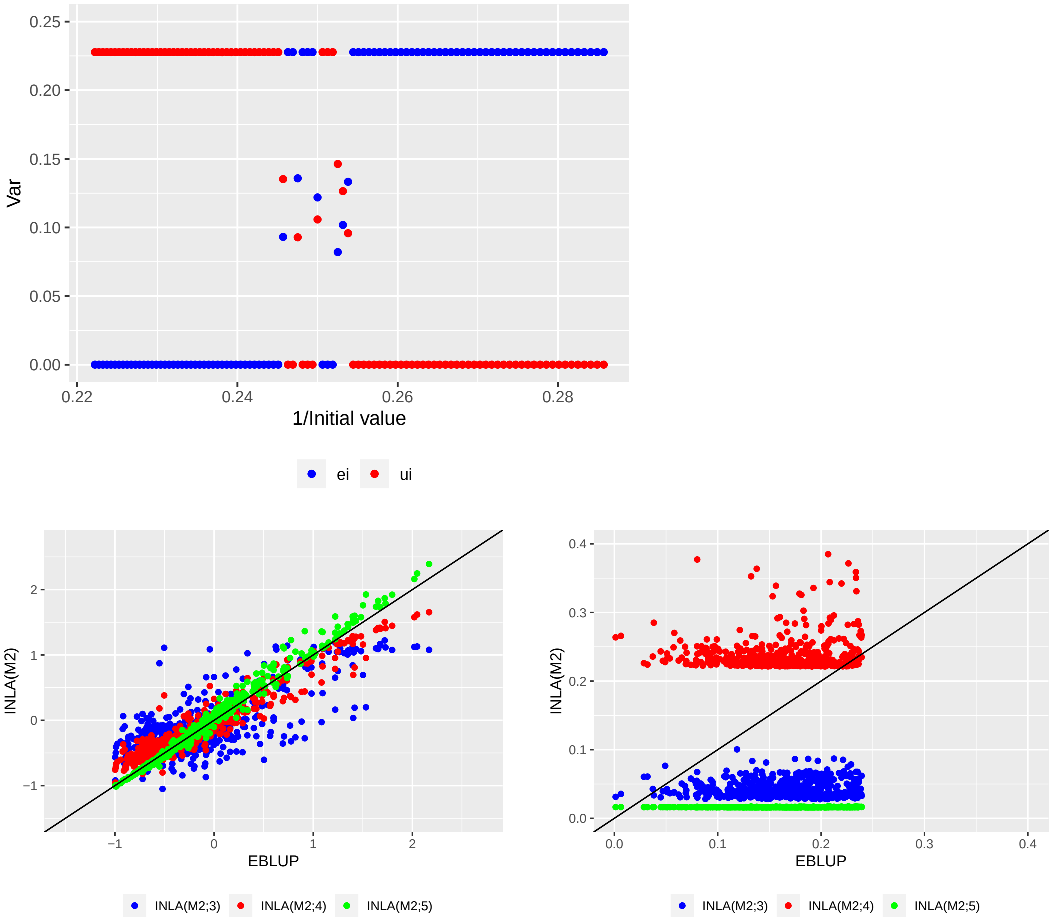


analyses that uses different initial values in the INLA algorithm. The top plot in Figure 2.c.4 plots on the y-axis estimates of the two variance components (in red and blue colour) against different starting values on the x-axis. We observe that depending on the choice of starting values, the INLA algorithm converges to different estimates of the variance components. The impact of converging to different estimates of the variance components on point and MSE estimates is illustrated in the second row of plots in Figure 2.c.4. INLA-based point estimates produced with three different starting values (represented by different colours) are positively correlated with EBLUPs but MSE estimates are clearly different.

2.c.4. Conclusions and further work

This section presents an application in which remote sensing covariates are used in area-level models for producing estimates of average wealth in areas in Bangladesh. The results from this small-scale application confirm previous studies i.e. that the use of remote sensing covariates can produce reasonable small area estimates. This is important because much of the effort in applications of small area estimation now focuses on approaches that rely less on Census data and instead use sources of auxiliary information that are frequently updated and are easily accessible. The application in this deliverable does not use mobile phone data although such data are available in the case of Bangladesh. The challenge with mobile phone data is that access requires special permission by the gatekeeper. Use of mobile phone data presents additional methodological challenges for example, relating to the definition of the small area geography of interest. In future work we plan to explore the use of such data and research relevant methodological issues. The present deliverable also reports some common pitfalls with the use of area-level models. Our analyses shows that care should be taken with model specification and assessing only point estimates may not be sufficient to conclude that the model has been well specified. In future work we plan to use existing methodology to account for the fact that

Figure 2.c.4.: Sensitivity analyses using different initial values in specification M2. Convergence of the variance components (top plot) and impact on points and MSE estimates relative to the EBLUP (bottom two plots).



the level 1 variance is estimated instead of being fixed. Additional work will include methods to account for the displacement of clusters affecting DHS data, the use of area-level models for estimating distributional parameters instead of means and proportions and the issues relating to the specification of models that include spatially correlated random effects.

2.d. Estimation of local poverty using remote sensing data

2.d.1. Introduction

One aim of the MAKSWELL project is the investigation of opportunities in big data. This section presents a case study application using remote sensing data to downscale poverty measures from official statistics by the University Trier (UT) and the Netherlands Statistical Office (CBS). Previous work packages have discussed findings and ideas from remote sensing in the fields of social statistics. Rather little activity was found for European countries compared to for example Africa (see Bejamin et al. (2017)), south America (see Burgess et al. (2012)) or Asia (see Xu et al. (2014), or Yue et al. (2014)). We believe this is because the European Union member states have established statistical institutions allowing for timely and reliable information for most parts of each country, specifically for bigger cities. This high quality of official statistics leaves less opportunities for advantageous use of remote sensing data.

Political decisions in cities and townships impact citizens' lives most directly. Important local level information, however, are often not available or suppressed for reasons of confidentiality. This hinders the possibility to investigate local structures of poverty. To successfully support evidence based political decisions also on such local levels this data gap has to be closed. In contrast, satellite data are globally available and confidential while being almost freely scalable. For this reason they might become an essential source of information for the investigation of local structures as used for estimation of GDP studied by Faisal et al. (2016), as well as global structures as attempted by Gosh et al. (2010).

Previous deliverables discussed possible measures and purposes of satellite based applications (see Deliverable 3.1, Chapter 4). To provide a contribution to the Makswell project under the project's timeline we decided to explore traditional, model based approaches. Although we expect that in the near future machine learning methods, such as adversarial neuronal networks and image recognition methods, will come to a greater focus in satellite assisted social statistics. Most such learning methods are extremely computer intensive, require labelled training data and high quality satellite images, which are mostly not freely available (see the application by Jean et al. (2016)).

The CBS and the University Trier cooperated to start an exploratory case study, investigating the opportunities of employing remote sensing data for disaggregating or downscaling official statistics on poverty and income. Downscaling, generally, refers to the derivation of fine resolution information from available information on a coarser resolution level (Zhang et al., 2014). The focus is on statistical downscaling methods that use the relationship between the target variable and the auxiliary information to predict unobserved small-scale information. For this purpose the CBS provided the median income in 500 meter statistical grid cells as well as auxiliary information on 100 meter grids for three mayor cities in the Netherlands: Rotterdam, Den Hague and Enschede. Due to confidentiality reasons, cells have been suppressed if less then 30 people live in the grid cell area. This results in a large number of missing values in the data set. More details about the data from CBS are provided in section 2.d.2.1.

Therefore, we exploit the opportunity of using indicators based on freely available remote sensing data as an alternative to the predictors provided by official statistics. Our concept relies on the assumption that people generally favour a certain composition of population density and urbanization in combination with access to leisure in a city and allocate along their preferences in combination with their income opportunities. This idea is plentifully discussed in economic research but usually falls short on opportunities to be evaluated empirically (for an extensive economic discussion of spatial equilibrium models see (Glaeser, 2008)).

To capture such ideas we initially use two indicators, the *Normalized Difference Vegetation Index* (*NDVI*), which assesses the density of vegetation, and the *Normalized Difference Building Index* (*NDBI*), which works as a measure of ground concealment or measure of built-up intensity (see Zah et al. (2010)). As a third indicator we combine both *NDVI* and *NDBI* into a single build-up (BU) indicator as presented by (Faisal et al., 2016) building on Zah et al. (2010). Those indicators are explained and described in detail in section 2.d.2.3.

Each of these indices is a rather simple combination of light bands and as such can easily fail to present the desired variable when the wrong combination of ground components come together as found by Xu et al. (2014) for night light images in China. Therefore, in section 2.d.2.2 we describe the CORINE dataset and how we use it to enrich the satellite indicators with additional information similar to the work of (Faisal et al., 2016) for Canadian cities.

Finally, in 2.d.4.2 we use these remote sensing-data based indicators for downscaling by first estimating a model on the larger scale of 500 m grid cells and, subsequently, predicting the local median poverty values. The study closes with a summarizing discussion of our results.

2.d.2. Spatial data

This section will present all the data we use starting with the data available from the CBS followed by explanation of the CORINE data and lastly of the satellite indicators and the indicators constructed from these. The following section is held small as a detailed discussion about all data available from CBS is provided in Deliverable 3.1, Chapter 4 of the Makswell project. Thereafter, we analyse the spatial structures in each city in section 2.d.2.3 to motivate the later models.

2.d.2.1. CBS data and grids

The CBS contributed two datasets. First, data about some poverty measures on a 500 × 500 meter grid, secondly some demographic, housing and service availability information on a 100 × 100 meter grid[1], for three cities. Some of the data are presented in more detail in section 2.d.2.3 together with the spatial distribution of the other data and indicators.

Although both data grids align on the edges and hence would be nested, they differ substantially. The poverty information is not available on a finer grid, to ensure the confidentiality of citizens. Grids with less than 30 households are suppressed for privacy reasons. To investigate factors of poverty, such differences in the data are a great problem. The auxiliary information are available for the 100 m grid as well as the 500 m grid. This section will present first the poverty data in detail and display differences, issues and comparisons of our data for the cities under study.

Poverty Data

Table 2.d.1 lists the poverty/well-being related information provided by the CBS on a 500 m grid cell level. The variables are mostly self explanatory and the CBS variable names will be used throughout this application. What has to be considered is that *medinc* is the only variable which is metric scaled, the other 4 variables are percentages. It is difficult to redistribute a percentage for local aggregation and disaggregation to different grid sizes. The median income hence was selected as relevant as well as easiest to handle variable and will be in the focus of our estimations in section 2.d.4.1.2.

Table 2.d.1.: CBS 500m Poverty Variables

Variable Name	Description
medinc	Median Income in the corresponding 500m grid cell
n	Number of households
plowinc1	Percentage of household below low income levels in 2017
plowinc4	Percentage of households below low income levels over a period of 4 years: 2014, 2015, 2016, 2017
psocinc1	Percentage of households below minimum social income levels in 2017
psocinc4	Percentage of households below minimum social income levels over a period of 4 years: 2014, 2015, 2016, 2017

[1] In what follows we call these grids 500 m grid and 100 m grid, respectively.

Auxiliary Information

Secondly, variables about each city is available on a 100 m grid size. The variables *HVAL*, containing the average value of houses according to the tax register, *OWN* the percentage of own houses contrary to rented house and last *ELEC*, the average electricity consumption per household showed specially valuable when estimation median incomes in the later models. These information are available both on a 500 m level and 100 m level, hence allow also for estimation on 100 m levels. Apart from these many other variables exist at the CBS, too many to state them here.

The Data Grid

Up to this moment it is not apparent why these datasets do not work well together. The following figures show the grid maps of each city, for 100 m and 500 m grids in violet and green colour respectively. The maps on the left hand side show the cities in full extend on top of the ESRI WorldTopo map for orientation. The right hand side figures are zoomed in on the city centre with HERE Hybrid Satellite images as a background.

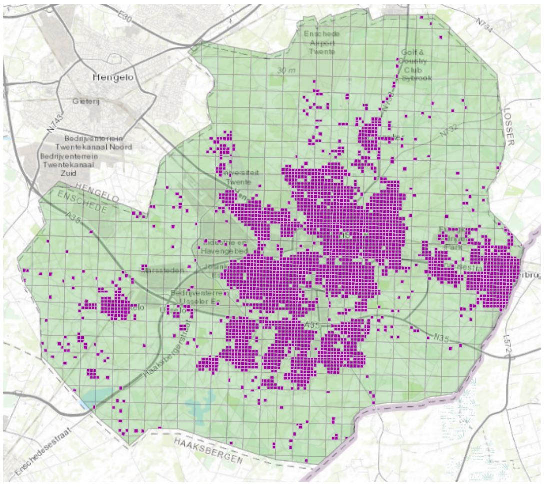


Figure 2.d.1.: Enschede grids on the *ESRI WorldTopo* maps background

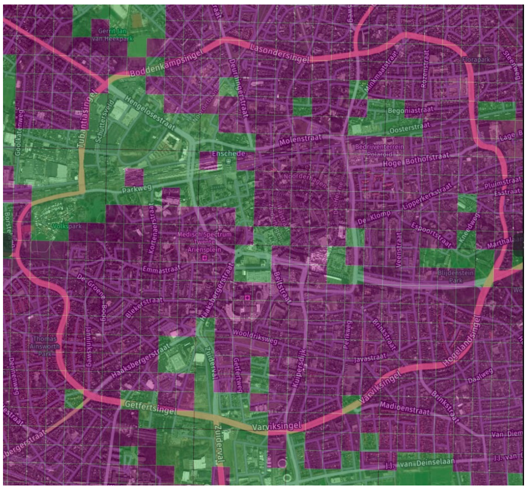
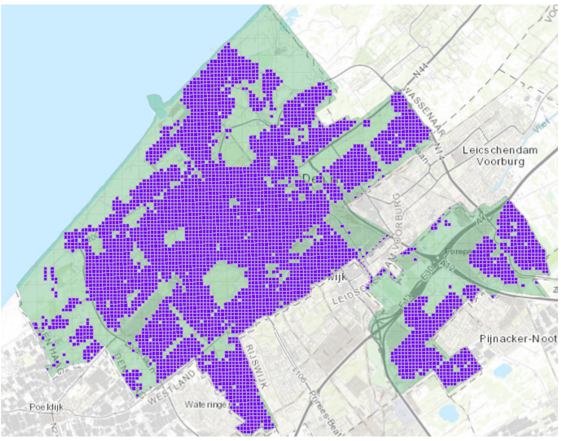


Figure 2.d.2.: Enschede grids on *HERE Hybrid Satellite* background

The grid map of Enschede (figures 2.d.1 and 2.d.2) is a compact map with nearly circular city boundaries on the 500 meter level. The 100 m grid looks very differently. The grid cells are clattered in bigger and smaller clusters around the entire map. The reason is that cells which would be placed on rivers and partially also industrial areas or greater streets do not exist. They are not classified as 0 population; as they naturally contain no households they are simple not accounted for. As a consequence, many island type small clusters exist, while residential settlements within the city are clearly visible. This leads the 500 m median income map in section 2.d.2.3 to contain many empty cells while the values in other 500 m cells are mostly provided by information which stem only from very few 100 m cells.

Den Hague, figures 2.d.3 and 2.d.4, is by far the most densely populated city. Only highways, beaches, parks and some industrial areas are non-populated, the core of the city, however, is almost uninterrupted continuous urban fabric.



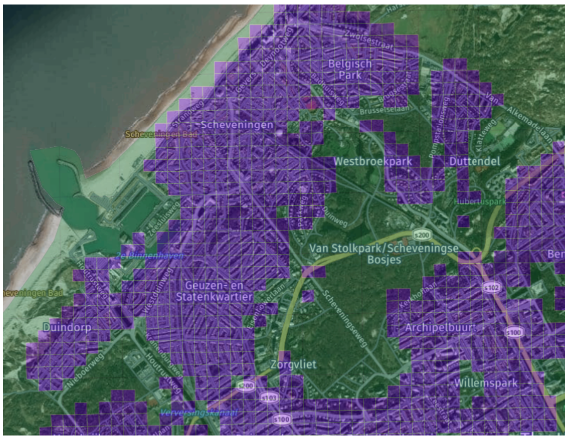


Figure 2.d.3.: Den Hague grid maps on *HERE Hybrid Satellite* background

Figure 2.d.4.: Den Hauge 100m on *ESRI WorldTopo* maps background

Den Hague is specifically separated in terms of its neighbours. To the south the city merges into Rijswijk, Voorburg and Delft. The city boarders are purely administrative but not physically visible for example by a river. We expect that analysis of cities like Den Hague will be specifically difficult at the borders, as with the data available we have to treat each city as an island and cannot consider the direct surrounding.

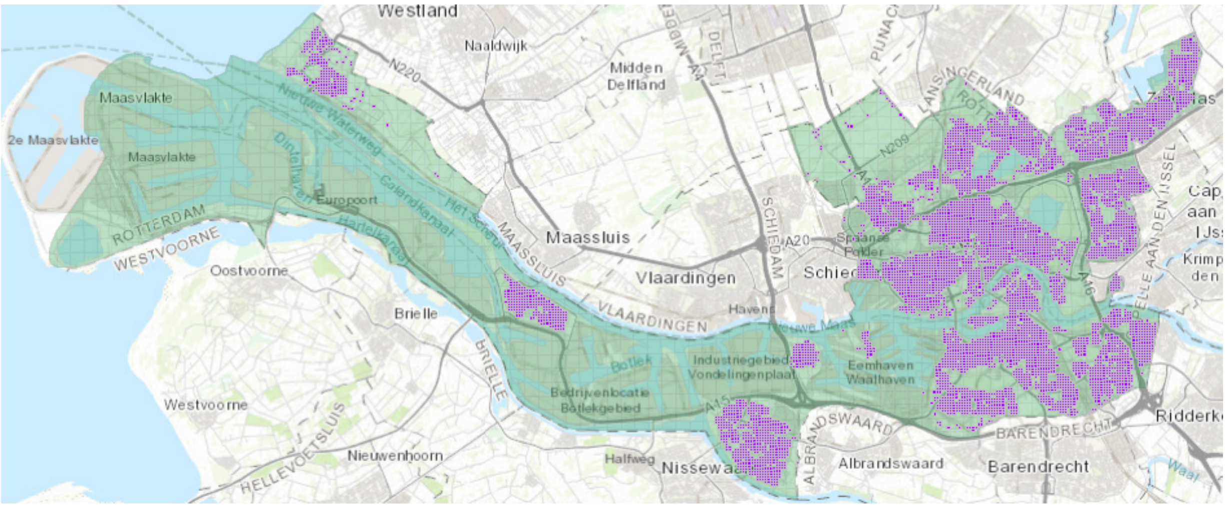


Figure 2.d.5.: Rotterdam 100m on *ESRI WorldTopo* maps background

Rotterdam in the figures 2.d.5 and 2.d.6 is far from circular compared to Enschede as the urban areas are distinctively spread out. Rotterdam consists of one greater, highly populated city centre to the east,

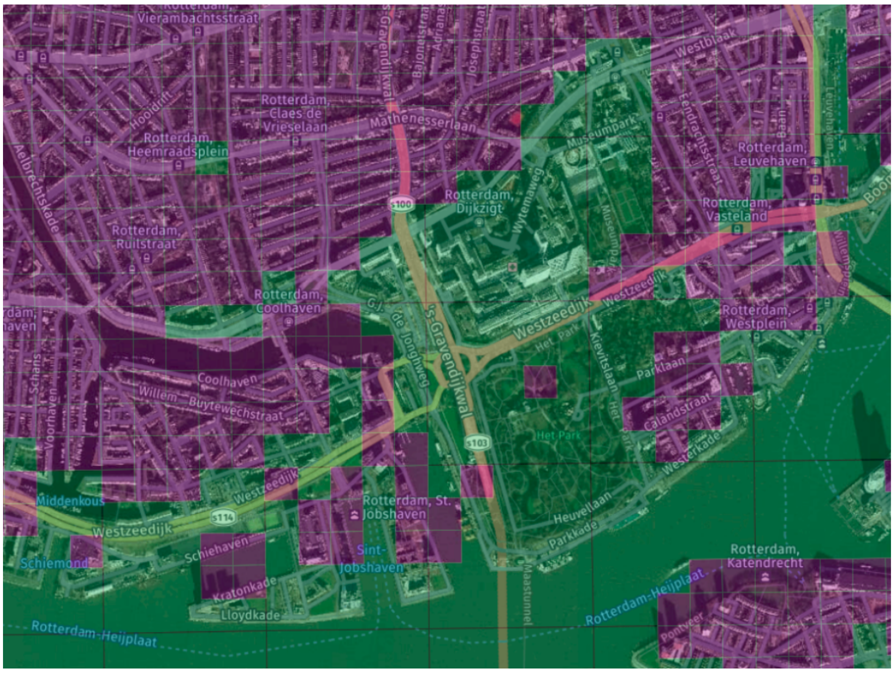


Figure 2.d.6.: Rotterdam grid maps on *HERE Hybrid Satellite* background

but reaches all the way to the ocean, because the port area belongs to Rotterdam. While Vlaardingen and Maassluis are separate cities, to the far west, smaller towns again are part of Rotterdam. Due to this, fact many grid cells are natural zeros. Rotterdam has by far the highest proportion of empty grid cells of all three cities. These cells will be ignored in the estimation as naturally empty area and suppressed/natural zero cells are differentiable. In figure 2.d.6 the city of Rotterdam is mostly compactly inhabited. The green areas stem from streets, waterways and industrial areas. Some of the riverside areas might contain almost only river, the centre however is populated densely enough to report even on such grid cells.

The 100 m grid cells contain no sensitive informations, hence cells are not suppressed. Unfortunately we do not know to what degree the suppression in the 500 m cells might be informative, but findings in section 2.d.2.3 suggest that depending on the city, high income citizens tend to live more remote. Suppression then would predominately effect high incomes, skewing the later model findings. This would lead to similar underrepresentation in our application as in many surveys on incomes. Hence all findings and discussions have to be treated under the assumption of non informative missingness and hence critically viewed.

The following section presents the CORINE dataset, the used satellite data and is followed by an spatial analysis of all variables.

2.d.2.2. Geo-Information-Systems

Geographic Information Systems (GIS) is a commonly used expression for spatial data with geographic reference information. This allows to project the information onto a map. Mostly such data are maps themselves.

Not every information possibly accessible by satellite images has to be deducted from satellite images directly. Many information about the composition and changes of cities are recorded and made available by the local administration.

While it can be difficult to get access to such information across cities and countries, the CORINE dataset is available for the entire European Union and beyond (38 countries). Updated every 6 years, information about the local land use from 1986 to 2018 is available. The *European Environmental Agency* (EEA) coordinates the CORINE project, but national data can be acquired from the corresponding national environmental agencies (see ESA (2020)).

The CORINE Land Cover Class inventory was initiated 1985 as a European wide harmonized collection of land cover information to support environmental policy, independent from national data centres. Europe is divided into 44 different classes of dominant land cover. Not every land cover class is present in each of the target cities, some are not present in any. Hence, we grouped the land cover classes to groups of urban, industrial, forest, agriculture and water area according to the table 2.d.2.

Table 2.d.2.: CORINE Landcover class grouping

Landcover group	Corine
Urban areas	0-1 : Artificial surfaces: continuous and discontinuous urban fabric
Industrial areas	2-10: Artificial surfaces: industrial commercial and transport units • Industrial areas • Roads, ports, • Mines, Dumps and construction sites • Non-agricultural vegetation areas (sport and leisure areas)
Agricultural areas	11-21: Agricultural areas: • arable land (irrigated and non-irrigated land) • permanent crops (wine, olives, trees) • pastures • heterogeneous agricultural areas
Forest areas	22-33: Forest and semi natural areas • all forest types • scrub and herbivorous vegetation associations • open spaces with little to no vegetation
Water areas	34-43: Water bodies and wetlands: • marshes and bogs • salt marshes and flats • inland water bodies • maritime water bodies

The idea behind using GIS is twofold. First by being able to eliminate for example water areas, we can correct the indicators which might misinterpret some land cover types such as water. On the NDBI indicator image for example, water bodies can appear systematically built-up which is not sensible in general. By contextualizing our data we can eliminate such measurement errors which might lead to miss conclusions. Secondly, information about vegetation or artificial ground concealment have different implications depending on the area they are placed in. Built-up spaces in an forest area might be playgrounds or parking areas and no indication of housing. By separating the land cover types we control for such dependencies between area type and the indicators. The next section presents the satellite data and indicators in more detail.

2.d.2.3. Remote sensing data and indicators

A great challenge in using remote sensing for social statistics is to find remote sensing data and derive indicator values which are meaningfully relatable to the variables of interest. As mentioned before, we want to generate indicators for vegetation density and concealment, which are derivable from the Landsat8 images. This section will present how the satellite data were compiled to generate the desired satellite indicators.

Landsat8 satellite images are public use data provided by joint efforts of *US American civil National Aeronautics and Space Administration* (NASA) and the *United States Geological Survey agency* (USGS) as part of the US National Land Imaging Program (NLI). Although requiring an account at USGS, anybody can get access to the USGS earth explorer under: `https://earthexplorer.usgs.gov/`. Earth explorer serves as a graphical user interface for multiple satellite data sources. As the satellite rotates around the globe the Landsat8 satellites take images with a ground cover of 190 km time 180 km in a specific succession. Each full scene of the Landsat 8 images finally is about 1.6 gigabytes in size (for more details see Department of the Interior, U.S. Geological Survey (2019)).

For the processing of the satellite images we used google earth engine (GEE), both for the spatial data compilation and as data catalogue. While the cities of interest were not problematically big, the use of GEE would allow also for applications to greater areas such as complete nations and more (see (Gorelick et al., 2017)).

USGS published the Landsat8 scenes in different qualities available under the term *Tiers*. Selected for use were the *USGS Landsat 8 Collection 1 Tier 1, real-time data raw scenes.* Highest quality, Tier 1 includes Level-1 Precision Terrain (L1TP) processed data. For those scenes information about radiometry and inter-calibration across the Landsat sensors were used to create time consistent images. USGS sets the quality tolerance for Tier1 scenes to $\leq$ 12 meters root mean square error (RMSE). The downside of Tier 1 images is the higher publication time. Raw scenes are published 14 days after recording, Tier 1 images require 26 days revision time before publication. (Department of the Interior, U.S. Geological Survey, 2019).

We also investigated the use of the *Sentinel-2 MSI Level-1C* images from ESA. The Sentinel-2 satellites are newer, publishing higher resolution images. While this should be advantageous, the standard algorithms did result in severe shadow issues in Rotterdam, which we could not solve for this project on time. Using Sentinel 2 images, would require cooperation with remote sensing experts to solve the specific issues we had with these images (European Space Agency, 2020).

Hence we finally used the Landsat8 data and used a common median filter for the image mosaicking and composition. This means that all the images from the Landsat 8 mission within a defined time frame are collected and combined. This creates a mosaic of images, partially overlapping each other in certain areas. These multiple impression of the same area are used to both create a final, homogenous image and to filter for shadows and clouds. In each area the median value of each band of the stack of images is selected as the composition value for the area of each band. This removes cloud impressions, with high values across all spectral bands, and shadows with little reflection across all spectra. That way one annual average image of all cities was created.

With the Landsat 8 images the *Normalized Difference Vegetation Index* (NDVI) and the *Normalized Difference Building Index* (NDBI) are calculated using simple ratios of spectral bands (see for example Faisal et al., 2016, p. 5 f.).

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{RED}}{\mathrm{NIR} + \mathrm{RED}} \tag{2.d.1}$$

The NDVI is the ratio of differences between the values of the near infra red spectral band (NIR) and the red light band (RED). Because plants reflect low amounts of red light, absorbing most of this spectrum for photosynthesis, a high NDVI value refers to dense vegetation. Naturally where a tree stands will be no skyscraper, but close proximity of vegetation and housing is possible and would influence the NDVI. On a 30 times 30 meters ground plot, houses and green spaces can exist next to each other, the resulting NDVI value would be in the mid range, not allowing separation housing and vegetation within one pixel (see Macarot and Statescu (2017)).

The NDBI works similar to the NDVI but uses different spectral bands to identify pixels with dense artificial surfaces, such as concrete or tarmac (see Zah et al. (2003)).

$$\mathrm{NDBI} = \frac{\mathrm{SWIR} - \mathrm{NIR}}{\mathrm{SWIR} + \mathrm{NIR}} \tag{2.d.2}$$

SWIR is the Landsat short wave infra-red light band. The NDBI value, exemplary for Rotterdam are plotted in figure 2.d.8 for comparison with a NDVI image in figure 2.d.7. For the NDBI, brighter spots mean more densely built-up areas, for the NDVI brighter pixel indicate more dense vegetation (see (Macarot and Statescu, 2017)).

This is visible in the lower left corner of figure 2.d.11, the white area is the *het Park*. The same area is slightly darker in the NDBI image.

However, the NDVI is sensitive to seasonal changes. The figures 2.d.9 and 2.d.10 show the difference in NDVI values from Den Hague (left) and for comparison Manaus, Brazil (right).

This demonstrates a problem with the NDVI indicator. Naturally in the Netherlands' climate many trees and bushes will loose their leaves which is the reason the NDVI can identify vegetation.

The blue spots indicate a negative difference, the NDVI value at winter is greater then at summer,

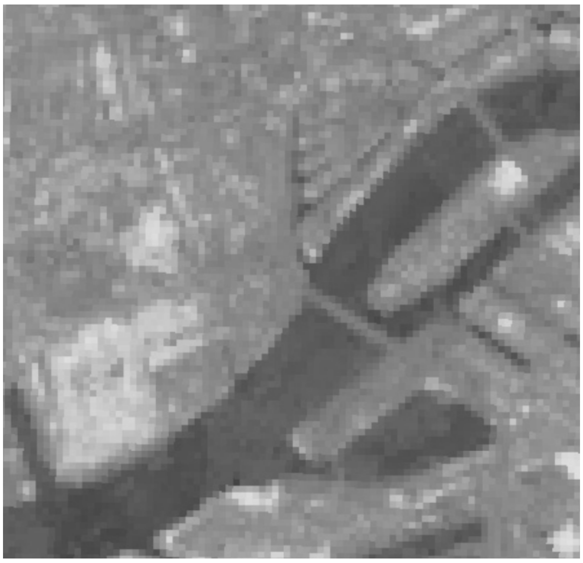


Figure 2.d.7.: NDVI index of Rotterdam based on Landsat 8 data from 2017

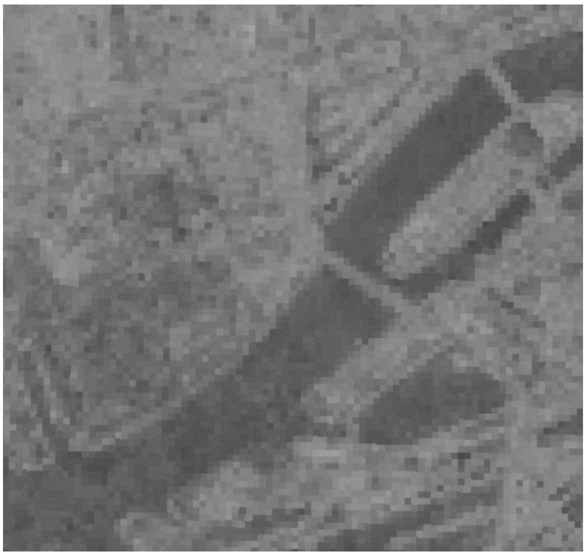


Figure 2.d.8.: NDBI index of Rotterdam based on Landsat 8 data from 2017

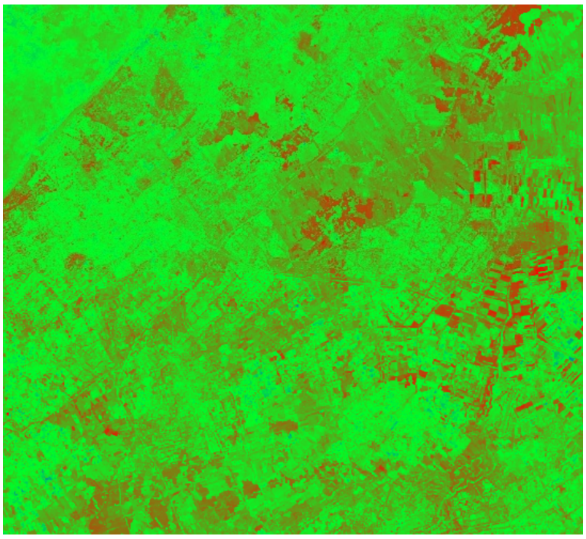


Figure 2.d.9.: Den Hague NDVI difference based on 2017 Landsat 8 images

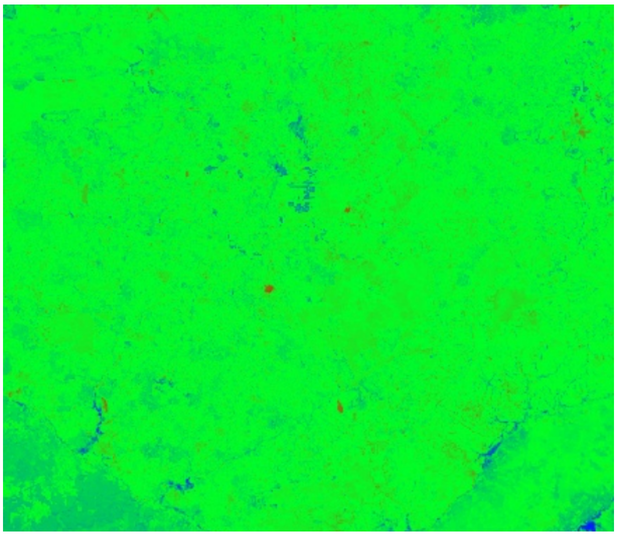


Figure 2.d.10.: Manaus NDVI difference based on 2017 Landsat 8 images

green means no difference and red spot show greater NDVI in summer compared to winter. Manaus is in Brazil in an evergreen forest area, meaning that there are no mayor seasonal changes. The image is mostly green with some few red and blue spots. The image from Den Hague shows much more deep red areas, indicating that in the time from spring to autumn the median NDVI values are much greater then during winter.

Although the median is already more outlier resilient than for example a mean, winter times in middle and northern Europa would mostly not contribute to knowledge about vegetation, but might skew the NDVI value distribution before applying the median filter. To avoid adding images without vegetation but with high proportions of cloud cover, only images taken between 01.03.2017 and 31.10.2017 were considered.

In combination, an area both scarce of vegetation, but dense on artificial surface is taken as area of intense construction, or an built-up (BU) area (see (Faisal et al., 2016)):

$$\text{BU} = \text{NDBI} - \text{NDVI} \tag{2.d.3}$$

Image 2.d.11 shows darker areas in parks and shaper edges at streets and housing compared to NDBI and NDVI. In contrast to Faisal et al. (2016), who classified the BU values binary, we did not find a suitable classification rule that applies to all the cities. We remain with the original BU indicators per pixel as they result from equation 2.d.3.

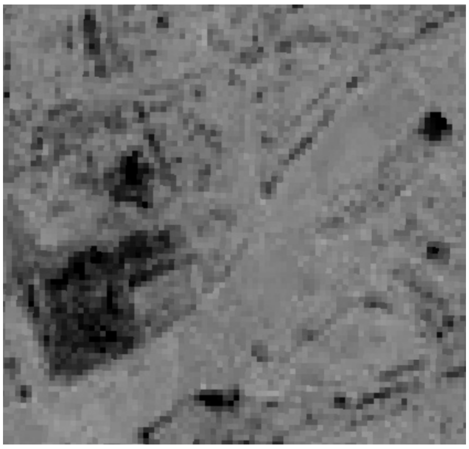


Figure 2.d.11.: BU-index map from Landsat 8 images of 2017

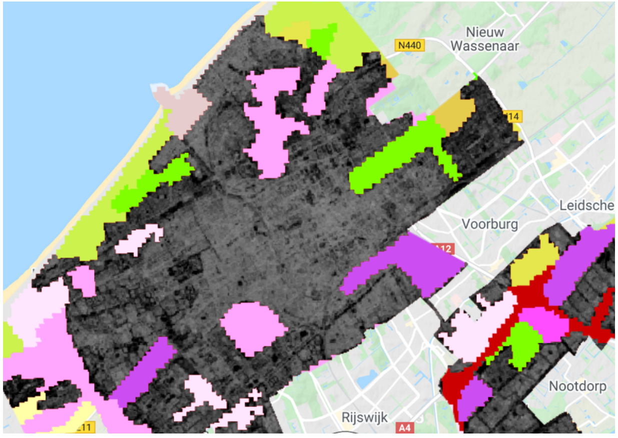


Figure 2.d.12.: BU Index and CORINE Map Den Hague

To decrease the sensitivity of some indicators to for example bodies of water we will be using the CORINE dataset to put the indicators in reference to the land cover class groups from section 2.d.2.2. Figure 2.d.12 shows again Den Hague, the grey scale parts are the BU pixel in urban areas, the coloured areas are different land cover classes which are removed for the urban area BU index of the city. We simply cut away all other CORINE class group areas from the indicator images of each city and then calculate statistics over the grid cell areas. How the indicators and official statistics data distribute in the cities is subject to the next section and will motivate the application of spatial and satellite based approaches in modelling poverty.

2.d.3. Spatial analysis

In this section we will present figures of spatial correlation and distribution of the median income in each of the target cities. While figures such as correlation do not constitute a real relationship, model based predictions will not work without any form of correlation between our target variable *median income* and any of the discussed satellite indicators.

To analyse the spatial distribution and dependencies of our variables we will mostly rely on the concepts of the *Moran's I* by Moran (1950) and *local indicators of spatial association* (LISA) as they were defined by Anselin (1995) as we are using the software *GeoDa* for the graphics and maps, which was developed under his lead. Before investigating the data in section 2.d.2.3 we shortly present how the Moran's I and the local Moran's I work.

2.d.3.1. Spatial statistics

Spatial analysis in form of the Moran's I and LISA statistics will be used to identify clusters of interest in each city and possible relations to those locations forming assumptions about the usability of the satellite indicators in modelling attempts. This is important to understand some specificities of the cities which will directly influence the possible success of modelling, as each city shows different challenges.

Moran's I The Moran's I statistics of spatial autocorrelation was proposed by Moran (1950), hence commonly used and further developed for example in form of G-statistics by Cliff and Ord (1981). Although Moran originally proposed his statistic, we will refer to the notation by Anselin (2018), whose software implementation *GeoDa* we use for the calculations and graphics.

In the focus of the Moran's I stands the relationship between an attribute x_i and it spatial lag x_j. So it is about the relation of for example the median income in area i and the median income of all the neighbours x_j. This is expressed by the deviation z_i of each area from the overall mean $\bar{x}$:

Finally, the Moran's I is characterized by Anselin (2018) as:

$$z_i = x_i - \bar{x} \tag{2.d.4}$$

$$\mathcal{I} = \frac{\sum_i \sum_j w_{ij} z_i z_j / (\sum_i \sum_j w_{ij})}{\sum_i z_i^2 / n} \tag{2.d.5}$$

w_{ij} is the so called $n \times n$ contiguity matrix containing the weights between all areas i and j for all $j \neq i$. n presents the number of areas. If the values in a neighbourhood are similar, z_i and z_j are nearly equal and a greater $\mathcal{I}$ results.

While this formula is similar to a scaled correlation formula, it is not used as descriptive statistics but rather as a test on the null hypothesis of no spatial correlation. By executing a permutation test, permuting the values for the areas j. The permutation results are used as empirical distribution function with which pseudo p-values are calculated by

$$p = \frac{R+1}{M+1}. \tag{2.d.6}$$

R is the number of times the calculated Moran's I for the permutations is higher than the actually calculated statistic from equation 2.d.3.1 and M number of permutations.

Weight Matrices

The spatial weights matrix determines which areas $j; j \neq i$ are considered as the neighbourhood of any area i and to which degree they are neighbours.

The definition of the spatial weights matrix w in equation 2.d.3.1 directly and greatly determines how many results of the spatial analysis look like. At the same time there is no common rule to determine the correct way of calculating spatial relationships. For this reason any such analysis must be take with caution.

As we are working with equally sized grid cells, the so called queen contiguity is going to be very similar to a distance weighted approach. We anticipate that the choice of a living place is directly affected by the immediate surrounding but also by areas further away, although less intensively. Therefore we decided for a distance weighted spatial weights matrix, accounting for the squared linear distance between the centroids of any area and all other areas in the city. In theory any area i can be a neighbour of every other area j to some degree, but this is not practical given areas far away will be assigned weights near zero but inflate the matrix x. The cities of interest are substantially different, specifically in their expansion so that we found no single cut-off distance which seemed well applicable to all three cities. For the cities Enschede and Rotterdam we found a great circle distance of 3 km appropriate, for the smaller Den Hague 2 km seemed suitable.

Local Indicators of Spatial Association

Standard modelling approaches of spatially referenced data would rely on the assumption of spatial stationarity of the relationship. Luc Anselin developed the *local indicator of spatial association (LISA)* as a concept to investigate this assumption in bigger, spatially referenced datasets. It will help us to identify local clusters and differences in the data, which will determine our modelling approach as well as help to understand special aspects which might be specific to each city.

These local approaches allow for deviations of areas from the global spatial relation. We will use the *local* Moran's I as measure of local association in relation to the Moran's I as measure of global association.

Anselin (1995)[p.94] defines the following two properties as defining for a LISA:

1. the LISA of each observation gives an indication of the extent of significant spatial clustering of similar values around that observation

2. the sum of LISAs for all observations is proportional to a global indication of spatial association

So the LISA should be hierarchical coherent such that a local Moran's I is just a spatial decomposition of the global Moran's I while allowing for a measure of significance.

To satisfy criterion 1 Anselin (1995) formulates a LISA as some function f over the observations y_i of an area i in relation to the J_i neighbour observations y_{J_i} :

$$L_i = f(y_i, y_{J_i}). \tag{2.d.7}$$

The neighbourhood J_i of area i is defined via a contiguity matrix w as before. To allow for a statement of significance regarding the cluster structure, Anselin (1995) operationalizes L_i:

$$Prob[L_i > \sigma_i] \leq \alpha_i, \tag{2.d.8}$$

with σ_i a critical value and α_i as a predefined pseudo significance level as result of a randomization test.

Condition 2, the consistency with global measures, implies a linear proportionality of the LISAs and their global measure:

$$\sum_i L_i = \gamma\Lambda. \tag{2.d.9}$$

Λ is the global indicator and γ some scaling factor.

As such a LISA is not a descriptive statistic but a test against the null hypothesis of *no spatial association* of the variable of interest. Anselin (1995) proposes a permutation approach to achieve pseudo p-values. The permutation character is achieved in regard to the neighbourhood. By resampling over the location, where y_i is held fixed, the observations of y in all other locations j are randomly reassigned. Anselin calculates an empirical distribution function as basis for a hypothesis test. In the following analysis and graphics we used 99999 permutations for the determination of the significance of the local Moran's I values against the global Moran's I. Condition 2 allows for the interpretation of outliers with common outlier definitions such as 1.5 times the interquartile range. For more details about this perspective see Anselin (1995) page 97 ff.

Local Moran's I

In (Anselin, 1995) suggested the local Moran's I as a special case of *Local Gamma* with

$$\mathcal{I} = z_i \sum_j w_{ij} z_j, \tag{2.d.10}$$

using *Moran's I* as measure of spatial association. z_i and z_j are deviations from the mean, summed over $j \in J_i$ with J_i the J neighbours of area i. w_{ij} is the contiguity matrix containing the weights about the degree of neighbourhoodship. In a first order queen contiguity all these weights are equal, in our application we used quadratic inverse distance weighted weights matrices in all cities, with varying cut-off distances. Hence the weights are different for each neighbour.

Figure 2.d.13 shows the grid map of Den Hague. The white grids are neighbours of the most centre grid. The dull green grids are not considered neighbours any more. As we employ an inverted squared distance based measure for the degree of neighbourhoodship, the connectivity map is nearly circular.

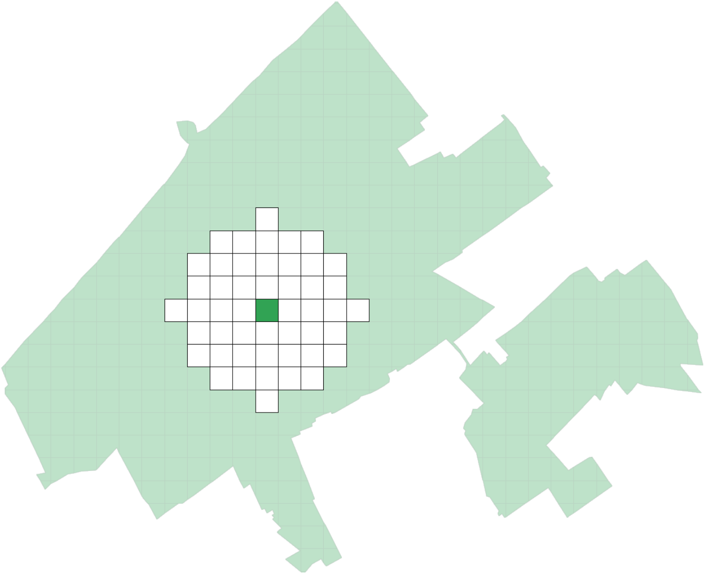


Figure 2.d.13.: Den Hague Connectivity Map

Anselin (1995) derived the first four moments of a randomization hypothesis of no spatial correlation based on Getis and Ord (1992) and Cliff and Ord (1981). By ensuring the global Moran's I to be the sum of the local Moran's I multiplied by some possible factor, the local Moran's I is simply interpreted in terms of deviation from the global Moran's I to identify outlying local areas, or clusters.

Finally the local Moran's I is given by the equation 2.d.11

$$\mathcal{I}_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{x}), \tag{2.d.11}$$

where i is the area i and x the associated attribute. $\bar{x}$ denotes the mean of the attribute. $w_{i,j}$ is still the spatial weight matrix between the different areas i and j. The variable S is used as a measure of variance according to the following equation:

$$S_i^2 = \frac{\sum_{j=1, j \neq 1}^{n} (x_i - \bar{x})^2}{n-1}. \tag{2.d.12}$$

In contrast to the G_i^* statistic by Getis and Ord (1992) high values of the local Moran's I indicate a cluster of similar values, rather than a cluster of high values. By accessing the position of the areas within the Moran's I scatter plot we can identify clusters of low and high values. We did however identify similar patterns with the local G^* statistic as with the local Moran's I.

It is important to recognise the test character of the local Moran's I, which is ignoring the multiple testing problem. Anselin (1995) mentioned that the use of Bonferroni (1936) bounds would be a possible way to account for the multiple testing problem, however it might also be too conservative implying individually significances of $\alpha_i = 0.0005$ to yield an overall $\alpha = 0.5$ when an application considers 100 areas. We usually consider even more area. Such the following analysis in section 2.d.3.2 will present cluster maps to analyse cluster tendencies of high and low values for median income and satellite indicators, next to a significance map to visualize the reliability of each areas classification assuming an maximal $\alpha = 5\%$ to just be still significant without the use of Bonferroni bounds.

We accept this shortcoming as we will discuss our later modelling compositions with observations won under the following spatial analysis and actually found results, such using the LISA statistic, but LISA is no primary model component.

2.d.3.2. Spatial clusters

We use the following figures and insights as starting point for the later models in the sections 2.d.4 and thereafter, and do not rely on the presented figures of spatial correlations alone. Here we investigate the population, the median income and some auxiliary structures. We identify structures and clusters that might give insights towards the applicability of satellite data based information.

Enschede

The figure 2.d.14 shows a quantile map of the population number on the left denoted as n and the median income *medinc* in the right hand figure 2.d.15, of the city Enschede. The dark grey areas are undefined as the data from the CBS did not contain information about the variable in the corresponding area. The population is mostly complete, but the median income map is to an overwhelming degree empty. CBS supressed any median income information if an area is inhabited by less than 5 people, which for the case of Enschede happens often.

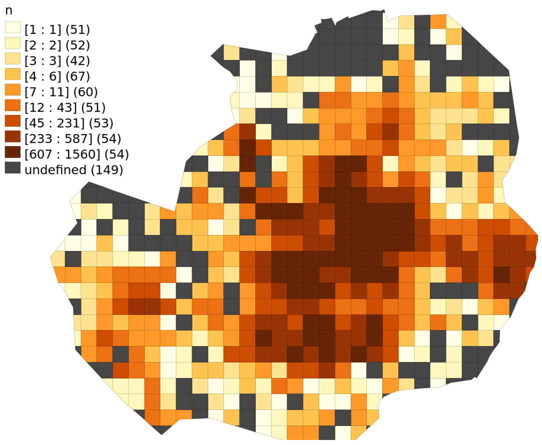


Figure 2.d.14.: Enschede Median Income LISA significance map, 3km threshold

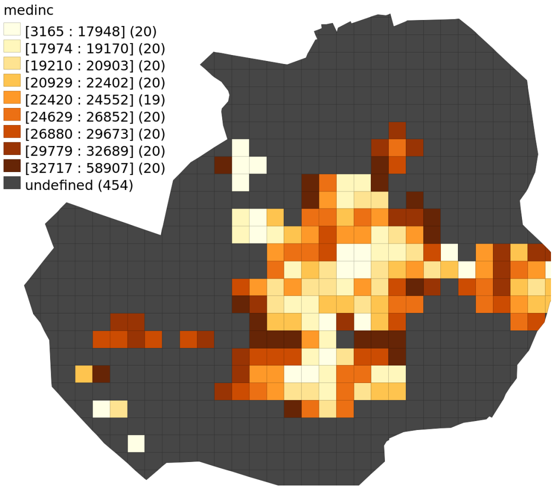


Figure 2.d.15.: Enschede Median Income LISA classification map, 3km threshold

The central part of the city around the district *Oude Markt* in the very centre is most densely populated and the population density is decreasing with rising distance from the central districts. An exception might be the district *Eekmatt*, which is directly located at the Netherlands-German border to the east and shows a high population density even though it is farther from the Enschede city centre.

Discussing the median income in figure 2.d.15 is critical, the high proportion of empty cells leaves little room for in-depth understanding. Of the parts we can analyse it seems that higher incomes live rather at the outside of the city centre and in the Netherlands-German border area.

Corresponding to the low amount of available information about the median income, a spatial analysis of median income is nearly meaningless. The Moran's I value for the median income is just 0.012, suggesting no relevant spatial clustering of median income in Enschede. For Enschede the suppression of income information leads to the immediate inability to use these data meaningfully to gather understanding of within city structures.

The figures 2.d.17 and 2.d.16 show the local Moran's I cluster and significance maps of Enschede. The legends show the meaning of the colour coding, in braces are the numbers of grids of the corresponding area. While the left hand figure shows all the classes without a minimum significance level the right side figure only shows the different significance for the categorization. While the classification maps suggests a cluster of low income in the city centre, the significance map however shows that spatial relations are not significant for almost all grids.

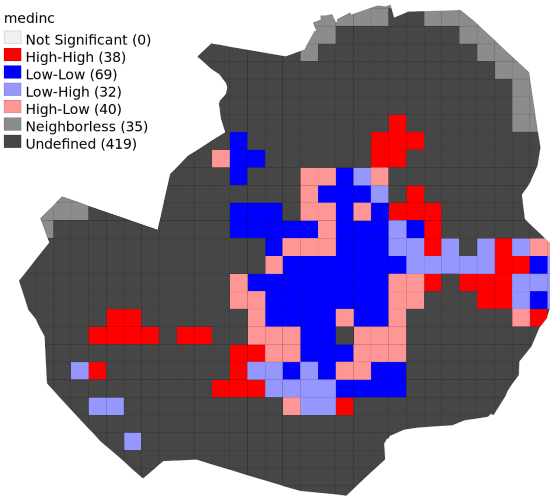


Figure 2.d.16.: Enschede Median Income LISA significance map, 3km threshold

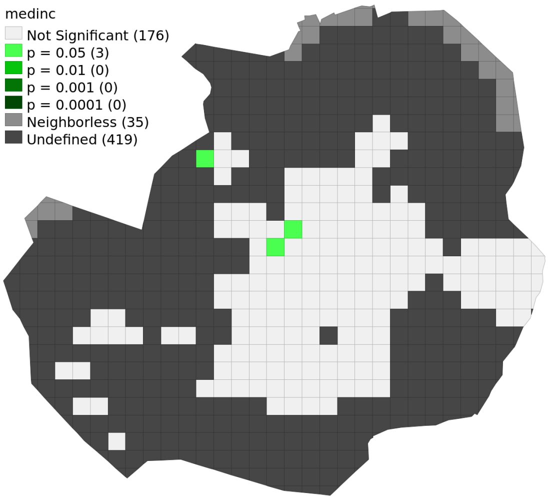


Figure 2.d.17.: Enschede Median Income LISA classification map, 3km threshold

Den Hague

Den Hague is much more densely populated, which allows for a more detailed view on the distribution of median income here. Compared to Enschede, Den Hague provides a different challenge. It is directly connected to other cities in the south and east, but neighbours the ocean to the north. The main part of the city is connected by a river section alone with *Ypenburg, Haagord* and *Leidscheveen* (see figure 2.d.18).

Just as in any of the other cities, Den Hague is mostly populated at its centre. Den Hague however is a city with strong tourism. Non urban areas in Den Hague are mostly park areas rather then agriculture or forests, the breaks in the population numbers are mostly due to parks as in the area around *Madestein* to the west and *Scheveningen* to the east. The non-defined areas are caused by the coast line and by the highway crossing between the A4 and A12 in the south.

The figure 2.d.18 shows the population number of Den Hague as a quantile map. In comparison to figure 2.d.19 it becomes visible that areas with particular low population density are hosting the most prosperous inhabitants mostly in the north eastern area in *Scheveningen*. Den Hagues' high income areas are allocated away from the city centre in lower population areas with direct access to green spaces and parks. The second group of high income earners allocate to the beach front in the north, under the condition to be not too close to the touristic peer area or the harbour.

Den Hague's median income has a Moran's I value of 0,218 for 0.00001 pseudo p-value signalling a substantial relation between the median income of an area and its neighbours. Figures 2.d.20 and 2.d.21 show the univariate local Moran's I of the median income. We can clearly see a

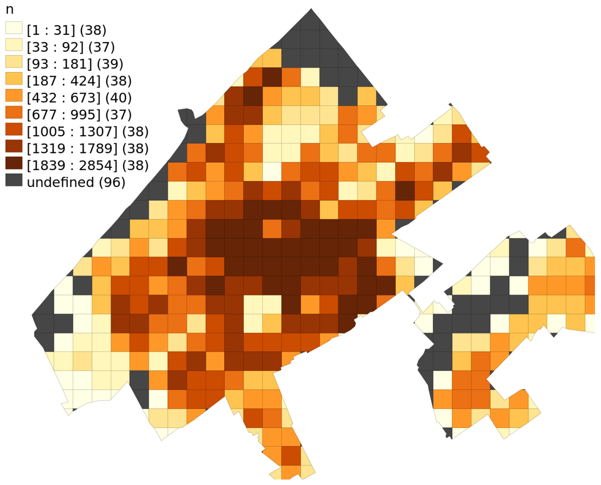


Figure 2.d.18.

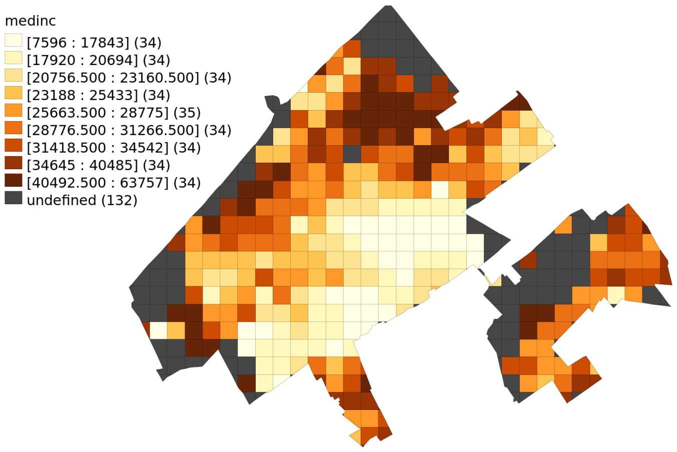


Figure 2.d.19.

separation between the wealthy northern area close to the ocean. The low income population appears almost polar opposite, located at the southern border. Another interesting cluster is located in *Kraayenstein* in the west city, where higher income is located directly at the Madestein Park in direct neighbourhood of low income areas just a bit further from the park.

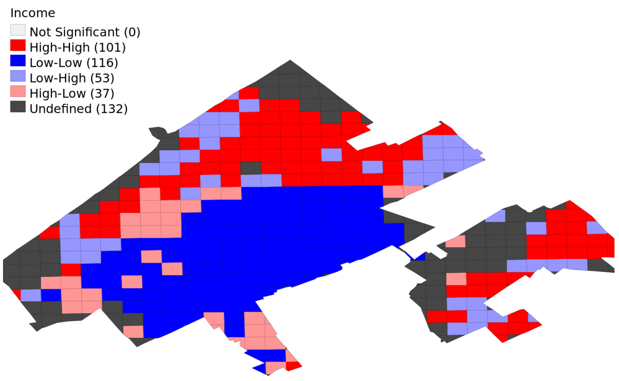


Figure 2.d.20.: Den Hague Median Income LISA classification map, 2km threshold

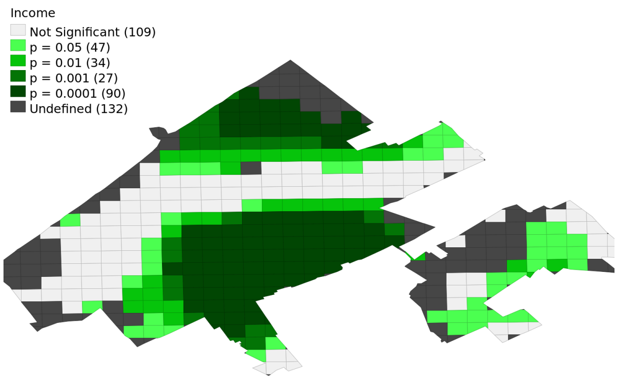


Figure 2.d.21.: Den Hague Median Income LISA significance map, 2km threshold

Please note that the cluster map 2.d.20 has no significant cut-off, but rather shows all the local Moran's I values. Figure 2.d.21 next to it shows areas with significant spatial clusters at least on a 5% level or lower as mentioned earlier, the significance has to be considered but critically reviewed as mentioned in section 2.d.3.1.

Rotterdam

Rotterdam looks completely different. The city is far from round shaped, and hosts 3 main population centres marked in the figure 2.d.22. For Rotterdam we are confronted with a particular problem that the overwhelming part of the city area has no relevant population count. Mostly this is because of the river leading from the Rotterdam harbour to the Northern See, areas which are mostly industrial compounds.

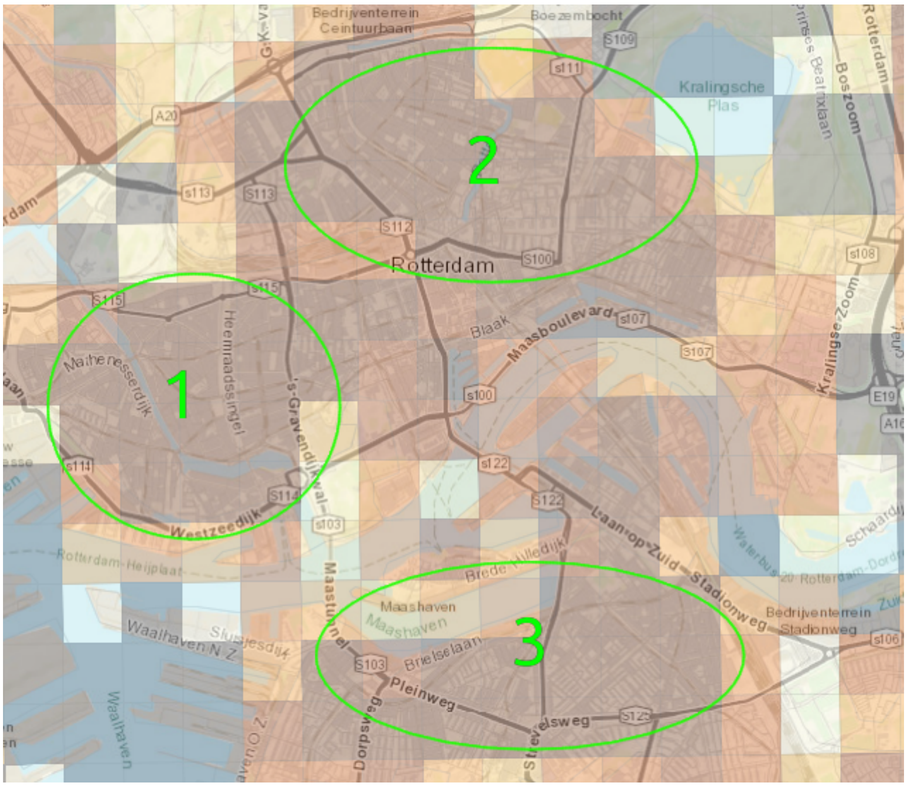


Figure 2.d.22.: Population Centres Rotterdam zoomed with ESRI WorldTopo Map as background.

In figure 2.d.22 the three densely populated areas in the city centre area separated north to south between areas 1,2 and 3 by the river *Nieuwe Maas* and west-to east between area 1 and 2 by the central train station. In fact Rotterdam is also rather circularly populated with the particularity that the population centre is broken up by infrastructure and waterways. The outer parts of the city though are scattered far further from the city then it has been the case in Enschede.

Figures 2.d.23 and 2.d.24 show the population and median income distribution, respectively.

Seeing the median income in comparison, we get a similar impression as in Den Hague. The most populated areas are also the areas with low income. The three densely population areas are also the areas with the lowest median income in the city, apart from very specific areas around Rotterdam Blaak. In tendency higher incomes can be found in the northern part of Rotterdam city, again near to parks, here *Kralings Plas* and *Bergse Voorplas*. This is interestingly a very touristic area in contrast to Den Hague where touristic areas were rather circumvented. The small towns *Rozenburg* and *Hook of Holland* to the east are rather mixed, but with a tendency to higher incomes.

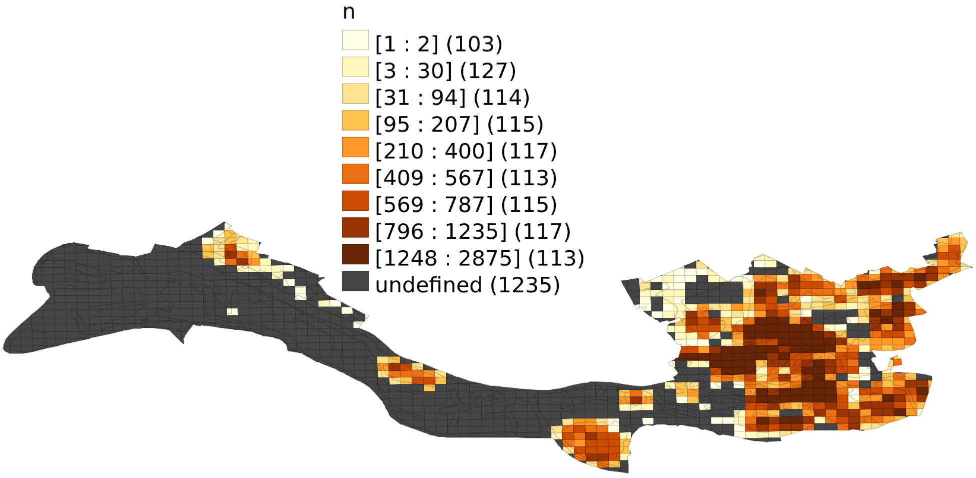


Figure 2.d.23.: Quantile Map of Rotterdam Population

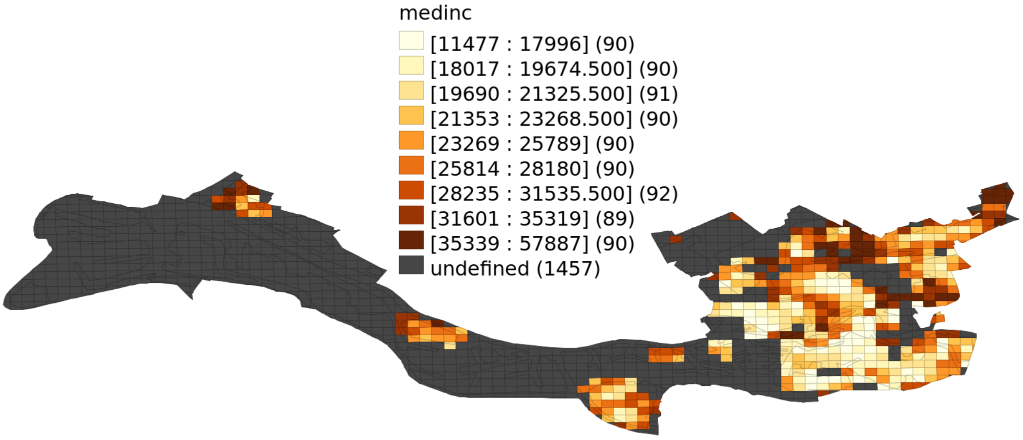


Figure 2.d.24.: Quantile Map of Rotterdam Median Income

The global Moran's I value for the median income in Rotterdam is 0.234 at a pseudo-significance value of below 0.00001 for 99999 permutations on the permutation test. The local Moran's I is presented in figure 2.d.25, again without a significance boundary for the classification, but with a significance map in figure 2.d.26 with a maximum value of 0.5. We employed a 3km distance threshold for the contiguity matrix.

Rotterdam is rather mixed, with high proportions of low income areas in light blue next to high income areas. Nevertheless high income areas seem to be clustered to the western part of the city, while lower income grids are mostly located to the west. Undefined areas again are placed mostly in parks, rivers and highway areas. Specific to Rotterdam is the high quantity of unidentified areas located around the river. These are harbours and expected to have little to no living population. Commonly these are not attractive living areas, leading to assume that in contrast to the other two cities the unidentified areas, if inhabited, might be predominately low income areas. Also in contrast to Den Hague, Rotterdam's high income areas appear to be in tourism attractive areas in the centre around *Rotterdam Nord* over *Rotterdam Blaak* to *Rotterdam Binnenhaven*.

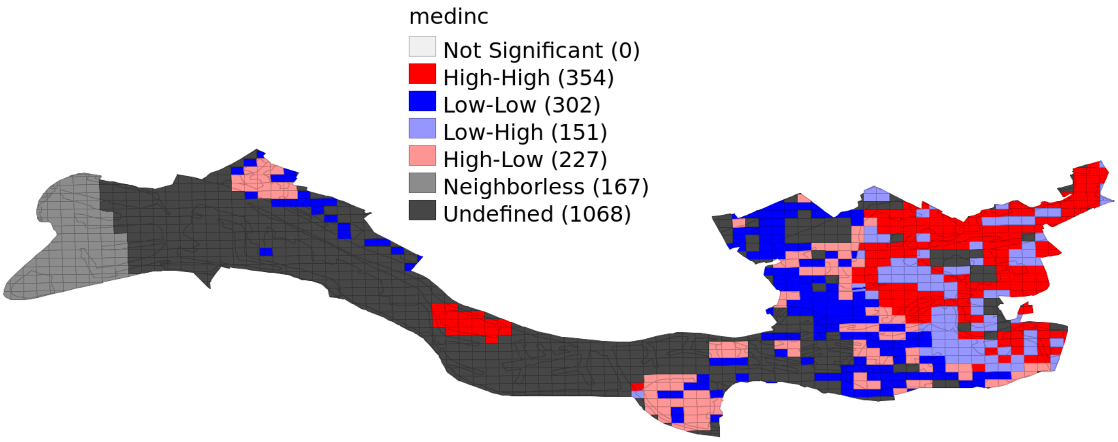


Figure 2.d.25.: Rotterdam Median Income LISA classification map, 3km threshold

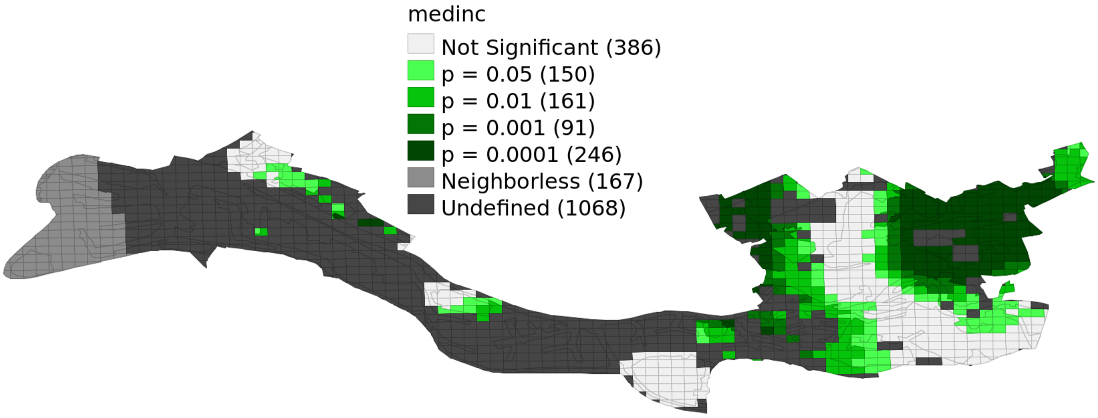


Figure 2.d.26.: Rotterdam Median Income LISA significance map, 3km threshold

Den Hagues observation would coincide with some discussions in spatial econometrics expecting a separating behaviour from richer people and a tendency for larger living spaces, even at the costs of potential travelling expenses, while lower income earners stick to highly populated areas such as city centres (see Glaeser (2008) for an economic discussion).

2.d.3.3. Satellite indicator relations

The analysis of median income in the previous section found that dense built-up is followed by low income, while access to leisure and green spaces is attractive to higher incomes. This leads us to suggest that population density, approximated by NDBI, and access to parks, approximated by NDVI might be good candidates of the estimation of poverty measures. This section presents some insights about the relation of each indicator with median income and later for the indicators in reference to the landcover groups.

Table 2.d.3 shows the correlation between the median income in the corresponding city and the median value for the NDVI as well as the sum of the NDVI over each grid cell. These correlations are calculated using only the information of grid cells for which median income was available without considering the location. The CBS data contain suppressed cells, which we did not considered here as they cannot be assumed zero. This will distort the correlation to some degree, as suppression takes place when too few people are living in an grid cell area. This naturally is the case in non-urban areas, leading to an artificial shortage of information in non urbanized areas. Hence the correlations should only be taken as indication for the potential usability of the satellite indicators concept rather then a direct evaluation of their performance.

Correlation	NDVI median	NDVI sum	NDBI median	NDBI sum	BU median	BU sum
's-Gravenhage	0.3609	0.3761	-0.14469	-0.1494	-0.3458	-0.3652
Enschede	0.4304	0.3294	-0.3275	-0.2178	-0.4180	-0.3057
Rotterdam	0.1779	0.08809	-0.3096	-0.2377	-0.2387	-0.16133

Table 2.d.3.: Correlations of Median Income and Satellite Indicators

The relationship direction is consistent in all cities, while the median value over the 500 meter grid cell pixel seems stronger related then the sum. Access to green spaces seem positively related with higher income. While this was particularly expected in Den Hague, see 2.d.3.2, the relationship seems strongest in Enschede. We assume that within the urban areas those with high income might be closer towards the outskirts of the city, increasing the proportion of vegetation, visible in the NDVI values.

Rotterdam has very little vegetation in the city, and it appears not to be a strong point of attraction for higher incomes, hence the NDVI is less relevant here. The degree of urbanization however seems rather strong related here.

A higher degree of urbanisation in form of the NDBI indicator is consistently negatively related to high median income. People with high income appear to avoid crowded, highly built-up, touristic or industrial areas of a city. NDBI is less correlated with income in Den Hague as in any of the other cities, however it is much stronger related with income in Rotterdam.

The BU indicator usually performs somewhere between the NDVI and the NDBI correlations. If uncertain which indicator to use, the BU appears rather stable over different cities. However, the NDVI and NDBI indicators might be used in the same model and possibly outperform the BU indicator in general. NDVI and NDBI should however not be both combined with the BU indicator in one model to avoid perfect multicollinearity.

We believe that using a multitude of indicators, or good combinations with additional data will be crucial to make models adjustable enough to account for city or area specific differences. Therefore we split up each indicator depending of the 5 described landuse groups using the CORINE dataset and revaluated the correlation between median income in an area and the NDVI, NDBI and BU statistics conditional to being in the corresponding landuse area.

If we differentiate between different land use types the correlation between the indicators and the median income changes to the values in table 2.d.4.

	Enschede			's-Gravenhage			Rotterdam		
	NDVI	NDBI	BU	NDVI	NDBI	BU	NDVI	NDBI	BU
urban	0.458	-0.039	-0.418	0.344	-0.3	-0.359	0.236	-0.187	-0.227
industrial	0.02	-0.18	-0.084	0.426	-0.2719	-0.399	0.2	-0171	0.204
agriculture	0.04	0.3051	-0.06	-0.08	0.187	0.139	-0.03	0.097	0.073
forest	0.15	-0.00014	-0.147	0.129	-0.35	-0.227	0.113	-0.217	-0.139
water	NA	-0.1766	NA	NA	NA	NA	0.159	-0.326	-0.27

Table 2.d.4.: Table of Landuse Group Specific Correlations of Indicators and Median Income

We can see some interesting differences between to the overall indicator comparison in table 2.d.3.2 and the differentiation by land use group in table 2.d.4.

In Enschede we can see that the NDVI is stronger positively correlated when only accounting urban areas rather than any area. This indicates that vegetation might indeed be relevant in general but specifically of interest in urban areas. It does not correlate with income in industrial or agricultural areas in Enschede. In reverse, the NDBI is mostly relevant in agricultural areas. Here a higher degree of urbanisation reflected in higher NDBI values, which simply means that given one is living in an rural area, people rather prefer a suburb or township with increasing income and don't like to live totally dislocated as we expected.

In Den Hague we originally expected to see the NDVI indicator to be important to cover the clustering of high incomes in park areas. While this was not so visible in table 2.d.2.3, differentiating by land use types brings up this idea again. Against the tendencies in Enschede or Rotterdam the NDVI is highly correlated with median income in Den Hague also in industrial zones. This seems counter intuitive, but in this application parks and similar areas are part of the group of industrial land uses. This would again support the idea that higher incomes stick to park areas in Den Hague, but not as particular in other cities.

As before, Rotterdam has rather low correlation values except for NDBI in water areas. While Enschede and Den Hague have almost no water area at all, in combination with possible area suppression a correlation was not computable for the most part in Enschede and Den Hague. In Rotterdam a city with massive harbours however, 500m grid cells with water and housing exist and are even very relevant living and business areas. Here high built-up is negatively related to median income in the harbour areas with direct waterfront. We believe these are areas in one of the most touristic parts of Rotterdam the *Maritiem District*. With high income these touristic parts are avoided as living place.

In summary of these analysis it seems that with higher income, crowded areas and places attractive for tourists are avoided. Depending on the cities this can mean very different types of areas, which need to be taken into account. This supports the believe that models most likely can not successfully be used across cities and areas for poverty measures, without crafting the satellite based models specific for the corresponding cities.

2.d.4. Models and estimation

The analysis of the cities leads us to believe that some information generatable by satellite information might be a suitable proxy measure for determinates of housing decisions in terms of the inhabitants income. We believe that the proximity of parks and leisure options increases housing prices, attracting dominantly higher incomes. We believe this to be measurable by indicators such as the NDVI. Namely for Enschede and Den Hague we expect the NDVI to be a relatively useful indicator, positively correlating with median incomes. Similarly the NDBI as a measure of artificial concealment will be negatively correlated with income, making it a differentiating measure in densely populated areas. Rotterdam did not show that easily differentiable patterns, but we observed strong cluster behaviour. We expect the remote sensing indicators to not show great abilities in allowing local poverty predictions. However, a spatial lag model might be applied here. Using GIS data such as the CORINE dataset will allow us to target the same variables more precisely. As parks were grouped with industrial areas, the NDVI and NDBI in industrial zones is expected relatively strong in Den Hague, while the urban areas will be relevant to all cities. While the NDVI is mostly relevant in urban areas, it has a greater meaning in industrial areas for Den Hague as well.

2.d.4.1. Spatial models

We believe that the attraction of a locations for high income earners is strongly determined not only by the composition in the own area, but also by the neighbourhood. This might be specifically true for the satellite indicators we consider. Normally it is not possible to live in parks directly, but close proximity might be attractive. Models, which consider only the own grid cell values will not be able to consider such relations.

This section will explore the suitability of spatial regression models for th median income estimation based on satellite data alone. Earlier results suggest that no satellite generated information can compete with strong indicator by the CBS such as the housing value. Therefore we shall focus on satellite indicators alone.

The spatial analysis of median income in section 2.d.3.2 suggests that income clusters geographically it can be assumed that a spatially lagged y model would work well. However, it would require the median income of all neighbours to be available. Such it might be suitable for imputation of some few areas, for estimation of many areas or for downscaling to 100 meter cells it is not suitable.

Instead we will focus on spatial models which consider a local relation between between median income and auxiliary variables in the neighbouring areas as well as a possible relation of the error terms.

The following section will first present possible models briefly, followed by the application to each city.

2.d.4.1.1. Local spatial models

All spatial models in their core relate to Tobler's first law of geography. According to Waldo Tobler: "everything is related to everything else, but near things are more related than distant things." (Tobler (1970) p.3).

Such, not only the NDVI of area i is relevant but also the NDVI values of the neighbours $j \neq i$ which might indicate that the neighbourhood is situated with gardens or that parks and green spaces are nearby.

The Manski model by Manski (993b) is a model of global spatial relation and considers that all variables in a regression model are related through space (see LeSage and Pace (2014)):

$$\mathbf{y_i} = \rho \mathcal{W}_{\mathbf{ij}} \mathbf{y_j} + \mathbf{X_i} \beta + \mathcal{W}_{\mathbf{ij}} \mathbf{X_j} \beta_{\mathbf{2}} + \mathbf{u_i} \text{ , for } \mathbf{u_i} = \lambda \mathcal{W}_{\mathbf{ij}} \mathbf{u_j} + \epsilon_{\mathbf{i}} \tag{2.d.13}$$

In the Manski model, the dependent variable of an area i is the result of the values in y in the neighbour areas j weighted by the weights matrix $\mathcal{W}$, the dependent variable X in i as well as by the dependent variables in all other areas X_j. The residuals of the Manski model compose of the weighted residuals of the neighbours in addition to the stochastic error term ϵ_i.

However, the Manski model is so complex that in our applications no solutions to the regressions exist. Within the Manski model many other spatial models area nested which have a proper solution. As we are interested in spatial x-variable relations and error-term relations three simpler models are of interest are our focus.

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathcal{W}_\mathbf{ij}\mathbf{X_j}\beta_\mathbf{2} + \mathbf{u_i} \text{ , for } \mathbf{u_i} = \lambda\mathcal{W}_\mathbf{ij}\mathbf{u_j} + \epsilon_\mathbf{i} \tag{2.d.14}$$

Assuming that the dependent variable is not spatially related: $\rho = 0$, the Manski model turns into the so called spatial Durbin error model (SDEM). Durbin (1960) proposed his approach for time series, but since the idea has been adapted to spatial problems for example in Anselin (1980). The SDEM allows for a spatial relation between y_i and neighbouring $X_{j \neq i}$ through the spatial weights matrix $\mathcal{W}$ as well as a spatial relation in the error term u, but not for y.

Assuming that $\beta_2 = 0$, the spatial error model (SEM) results (see LeSage and Pace (2014) p.5):

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{u_i} \text{ , for } \mathbf{u_i} = \lambda\mathcal{W}_\mathbf{ij}\mathbf{u_j} + \epsilon_\mathbf{i} \tag{2.d.15}$$

The residual u in 2.d.15 is a function of the neighbouring residual values as well as a stochastic error term. This would account for missing of spatial related missing variables through the $\lambda W u$.

If one instead assumes $\lambda = 0$ a spatial lagged X (SLX) model results, which only assumes a spatial relation in the independent variables (see LeSage and Pace (2014) p.5):

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathcal{W}_\mathbf{ij}\mathbf{X_j}\theta + \epsilon_i \tag{2.d.16}$$

In our applications we used the Akaike Information criterion (Akaike, 1973) for initial variable selection to select a starting model then executed Lagrange Multiplier tests to compare the spatial alternatives as suggested by Anselin (1988).

2.d.4.1.2. Results of spatial models

This section is designed as a excursion to investigate of spatial model options might be an important model choice for satellite based indicators. However we were not able to implement everything that is discussed in other model sections. Hence, the application is limited. We will here only consider the direct satellite indicators, but not the indicators from combining satellite data with the CORINE dataset. Similarly an estimation on the 100 meter level was not possible in time.

For every satellite indicator, a group of statistics has been calculated such as the minimum and maximum values within the 500 meter grid, the variance median and mean. We started with a satellite only based OLS model, containing all available indicators. We applied the Akaike criterion for model selection. The result is the baseline satellite model for the investigation of spatial component and the possible advantage of spatial models in our application. This purely data driven approach as we have no experience helping us to make a knowledge base selection.

Den Hague

Den Hague turned out to be the city profiting the most from an spatial modelling attempt. Few NA areas and a compact city allows to actually relate many areas to each other. Only 37 areas has been suppressed which we will estimate here based on spatial models and compared the impact on the median income distribution in each city.

The application of the Akaike criterion for stepwise both directional variable selection leads to a model containing the following variables:

- BU indicator mean
- BU indicator variance
- NDBI indicator variance
- NDVI indicator minimal value
- NDVI indicator variance

Each of these variables are statistics over the pixel of the satellite images within the 500 meter grid area. Although not initially anticipated, also statistics such as the variance or the minimal value seem to be interesting information apart from median or mean statistics. The NDVI minimal value will give us information about whether there is at least on 30 by 30 meter area in the statistical grid, which contains vegetation, which might be more relevant to the citizens then the average degree of vegetation. Similarly, the variance might indicate how heterogeneity the area is, such whether vegetation and building exist in proximity to each other, which would be the case for greater variance values. In this constellation, the model is only able explain about 25.5% of the variation in the median income in Den Hague.

We execute a Lagrange Multiplier (LM) test on whether any of the mentioned spatial model options might be preferable against the simple OLS approach. In fact the LM diagnostic suggests that any form of spatial model would improve on the OLS approach with a p-value between 5.559e-06 for a robust spatial lag y model and 2.2e-16 for a spatial error model. After fitting all mentioned alternatives we compare the goodness of fit by pairwise comparison using a Likelihood ratio test for the lag x and error models from previous section, hence they are nested.

On a 5% significance level the likelihood ratio test suggests that the Spatial Durbin Error model as the most complex of options of applicable spatial models might be the best choice. Such we are able to anticipate that not only the satellite information at a location is relevant for the median income of it inhabitants, but also the surrounding area is relevant. Using a Spatial Durbin Error Model allows us increase the R^2 statistic for Den Hague from 25.5% to 45.3%.

If we consider auxiliary variable from the CBS, specifically the housing value information in variable "WOZWONING" the situation changes. Apparently the housing prices already account for the neighbourhood to such a degree that a spatial modelling approach would not improve on the OLS baseline models.

37 areas in Den Hague were suppressed for confidentiality. We estimated their expected values based on the presented spatial satellite model. The median income of the predicted areas is 32111 Euros, which is slightly higher then the 27308 in the unsuppressed areas. After model imputing the suppressed areas, the overall median income raises slightly from 27308 to 27792 Euros.

Enschede

Enschede was a difficult city in the entire application specially because so many grids in the city are uninhabited or suppressed that only few data remain.

Following the same model selection as in Den Hague, only the variables:

- BU indicator minimal value
- BU indicator maximum value
- NDVI indicator minimum value

This model only allows to explain about 20% of the present variation in the median income in Enschede. Again, using the CBS auxiliary variables in addition to satellite data a total R^2 of 0.8 can be achieved. The LM test diagnostic suggest that the spatial error model might be an improvement barely not significant with a p-value of 0.6. The Likelihood ratio test also suggests that no spatial model is an improvement over the the simple OLS solution.

In Enschede we did not find fitting indicators for a strong relationship with median income. Either no spatial relation exists, meaning that the median income in Enschede is rather randomly distributed in the city, or given the missingness issue in the city we were not able to identify such on a relevant significance level. Hence we will no further present the predictions for Enschede here.

Rotterdam

Rotterdam has 119 suppressed grid cells. If we follow the exact same approach again in Rotterdam. The following variables are selected:

- BU indicator sum
- BU indicator mean
- BU indicator maximum
- BU indicator variance
- NDBI indicator sum
- NDBI indicator mean
- NDBI indicator maximum
- NDVI indicator minimum
- NDVI indicator maximum
- NDVI indicator variance

This model achieves an R^2 value of only 23.7%. However, contrary to Enschede, the LM test suggests that we have strong spatial component we want to consider by using a Spatial Durbin Error Model as in Den Hague. Using a Spatial Durbin Error model does however only improve the model fit from 23.7% to 32%. Even with lower R^2 values the model imputation is able to estimate median income variables ranging between 11035 and 48879 Euro in the former NA areas alone. These spatial models might be suitable for a combination for synthetic downscaling methods, which did not allow for much deviation in the local estimates. Due to the spatial model imputation the median value of the median income changes from 23901 to 24932. Just the suppression led to an underestimation of the cities median income by 1000 Euros. This implies, given the model was true that indeed higher income areas are more commonly suppressed. But one imputed area is also a new lowest income area in our data with 11035 euro median income against 11477 euro before imputation.

Conclusion of Spatial Models

It would be interesting to further investigate how spatial weight matrices might be important for satellite based models by investigating the 100 meter grid cells as well. Unfortunately this was not possible here any more. The results presented however lead us to believe that satellite data as recordings of specific places cannot consider spatial relation which are relevant to social phenomena. Therefore, spatial models might be an important model component for satellite applications.

Specially Den Hague showed us that spatial models can be an important modelling approach to consider when using remote sensing information, while it might not be as helpful when using official statistics variables. If we consider strong auxiliary variables from the CBS such ans *WOZWONING* this tendency towards spatial models does not exist any more. The LM test would not support the use of a more complex model. The variable *WOZWONING*, the value of housing does already take the surrounding into account. For an application which relies on remote sensing information alone, using the possibility of spatial model seem feasible and necessary, while it might not be as relevant using other data from official statistics.

The downside of spatial models in the context of this application is that we cannot consider all cities together. The spatial models require "non-island" status over the data. This means we cannot accept and area which has no neighbours. As the cities area island for each other we cannot estimate the jointly, unless we change the application such that the entire Netherlands are part of the model.

2.d.4.2. Downscaling the median income using remote sensing data

2.d.4.2.1. Problem setting

We are now interested in the usability of the remote sensing indices for obtaining unreported income information on the very fine resolution level of 100 m grid cells for the cities under investigation. As stated above, information on the median income is only available on a resolution level of 500 m grid cells. The statistic is based on register data. In cells with less than 5 observations the information is suppressed. Further, we have auxiliary information from administrative sources for 100 m grid cells nested in the larger cells mentioned above. Again, information is suppressed in all cells with less than 5 valid observations. This results in a large number of NA-values in the auxiliary information. Contrary to the data from official statistics, the indices from remote sensing data described above are available for all cells under study.

The aim is to disaggregate or downscale the median income from the level of 500 m grid cells to the finer resolution level of 100 m grid cells with statistical methods. Generally, downscaling refers to the derivation of fine-scale information from available information on a coarser resolution level, i.e. to the transfer of data from a coarser to a finer scale (Zhang et al., 2014). As such, it finds broad applications in geoscientific research, especially in disciplines such as climate science, meteorology and remote sensing (see Zhang et al. (2014), Sec. 1.3 for an overview). In the context of social statistics, it might become relevant if for example otherwise unavail-

	Enschede	Rotterdam	Den Haag
value of houses (HVAL)	0.87	0.84	0.9
percentage of own houses (OWNH)	0.81	0.78	0.76
average electricity consumption per house (ELEC)	0.69	0.75	0.79
social security benefits (SOCBEN)	-0.49	-0.57	-0.62
average household size (HHSIZE)	0.7	0.56	0.43
perc. of people with dutch nationality (DNAT)	0.54	0.62	0.56

Table 2.d.5.: Overview of auxiliary information and correlation with Median income (500 × m grid cells)

able low-scale information is required as input for a statistical model or a composite indicator, combining range of indicators which might be available on different resolution levels (see Articus et al. (ming) for an example). A range of methods, which can broadly be categorized in statistical and dynamic approaches, has been discussed in geoscientific research. We focus on statistical approaches, i.e. on approaches that use statistical relationships between observed variables to predict unobserved small-scale information.

Obviously, statistical downscaling crucially depends on the availability of auxiliary information, which is generally hard to obtain on the very low target level considered. For this study, the CBS provided auxiliary information from administrative sources. The data contain valuable information and can generally be considered as good predictors. However, on the fine resolution levels of 100 m grid cells there is a large amount of missing data due to confidentiality reasons. Thus, if only administrative data is employed, for a large number of less densely populated cells no estimate can be obtained. The remote sensing data products decribed above might,therefore, be a valuable source of information in statistical downscaling with a very fine-resolution target level.

2.d.4.2.2. Disaggregating the median income

As stated above, the aim is to downscale the median income for the cities of Enschede, den Haag and Rotterdam from the coarser resolution level of 500 m grid cells to the finer resolution level of 100 m grid cells.

Table 2.d.5 gives an overview of indicators from administrative sources provided by CBS and their correlation with the median income for the three cities in the study. It can be taken from this overview that a range of highly correlated auxiliary variables for statistical downscaling is available.

Each of this indicators is available both on level of 500 m and 100 m grid cells. Information in cells with less than 5 observations is suppressed due to confidentiality reasons. Especially on the finer resolution level this results in a large number of empty cells. See Table 2.d.6 for the share of NA-cells in selected covariates.

	Enschede	Rotterdam	Den Haag
value of houses (HVAL)	0.43	0.18	0.14
percentage of own houses (OWNH)	0.57	0.39	0.32
social security benefits (SOCBEN)	0.68	0.45	0.44
perc. of people with dutch nationality (DNAT)	0.38	0.14	0.12

Table 2.d.6.: Share of NA-cells per variable and city (100 × 100 m resolution level)

A first, naive approach to obtain a disaggregated statistic is to set the mean value (or proportion) of the smaller cell to the value of the larger cell it is nested in, i.e.

$$\overline{Y}_i = \overline{Y}, \quad i = 1, \dots, m, \tag{2.d.17}$$

where $\overline{Y}$ is the larger area statistic and $\overline{Y}_i$ is the respective value for the smaller cell $i, i = 1, \dots, m$. We call this approach a synthetic downscaling approach. Obviously, it relies on the assumption, that smaller cells nested in the larger cell are homogeneous and share the characteristic of the larger cell. It is inappropriate if this very strong assumption is violated. Note however, that (given all smaller cells nested in the larger cells are taken into account) it ensures that central features of the distribution of the variable of interest like extrema, quantiles and the mean are retained. It is thus considered as a benchmark against which other approaches have to compete. Respective downscaling results for the median income are depicted in the maps in Figures 2.d.27 to 2.d.29.[2] As expected, the block-structure of the larger grid cells is clearly visible.

[2] All maps of downscaling results are produced with the open-source software GeoDa (Anselin et al., 2006) using the HERE light map as background map.

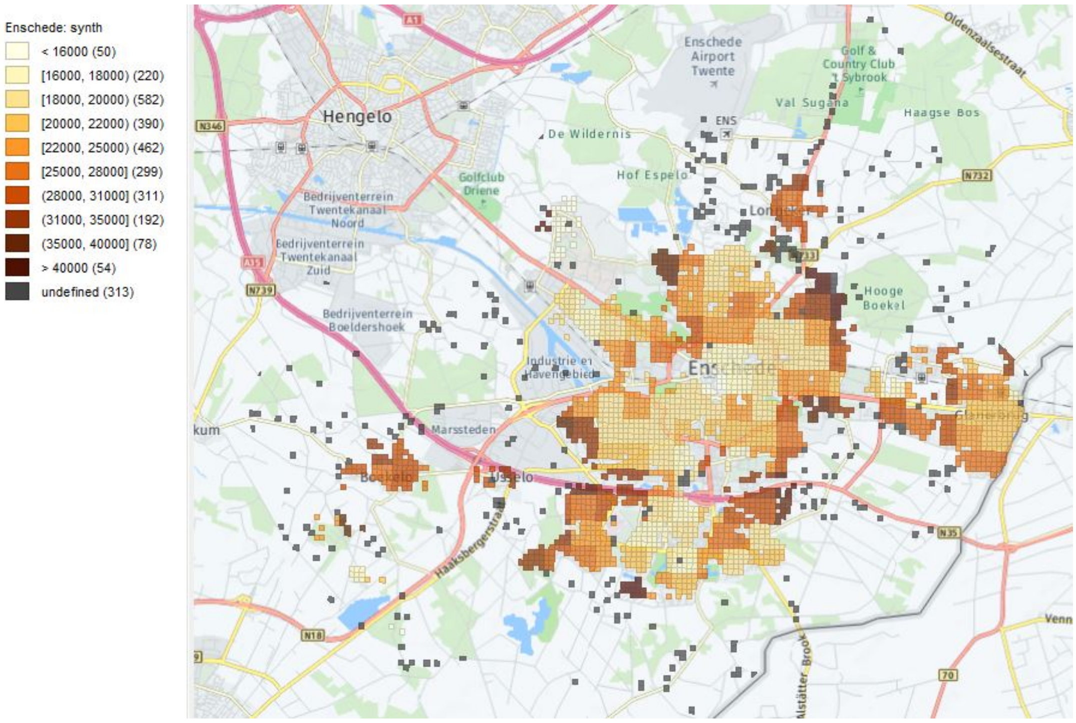


Figure 2.d.27.: Downscaling-result for Enschede
Synthetic approach

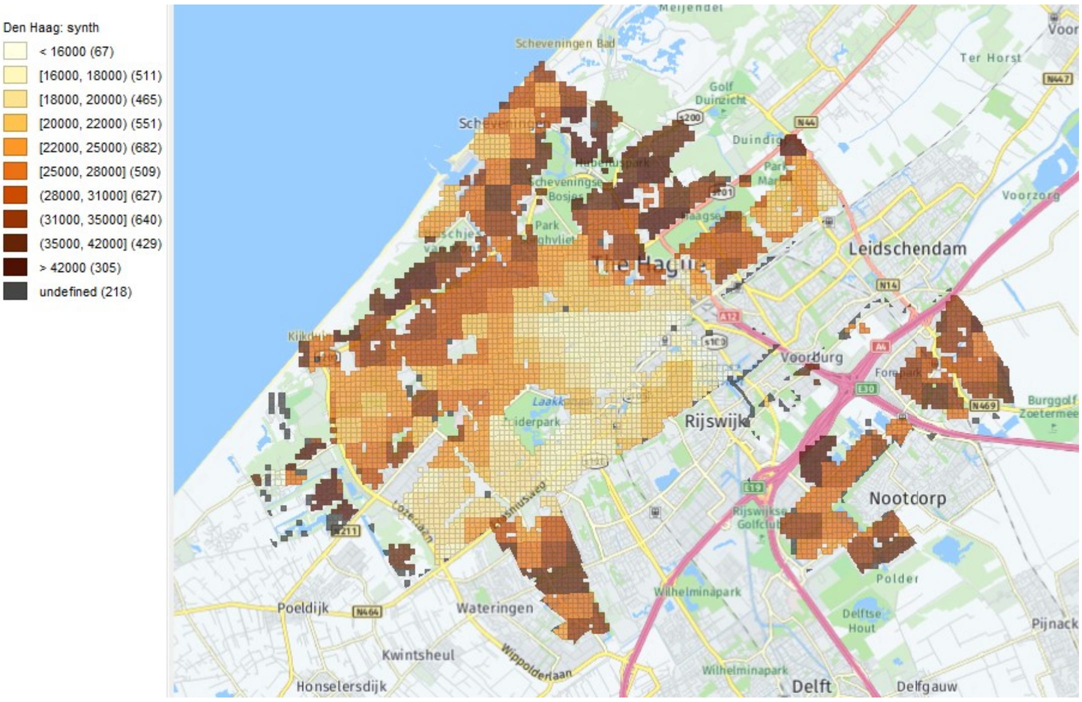


Figure 2.d.28.: Downscaling-result for Den Haag
Synthetic approach

	Estimate	Std. Error
(Intercept)	13298.78***	418.55
HVAL	40.46***	0.99
OWNH	133.03***	5.50
ELEC	−0.82***	0.22
R^2	0.87	
Adj. R^2	0.87	
Num. obs.	895	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2.d.7.: Model 1: Auxiliary information from administrative sources

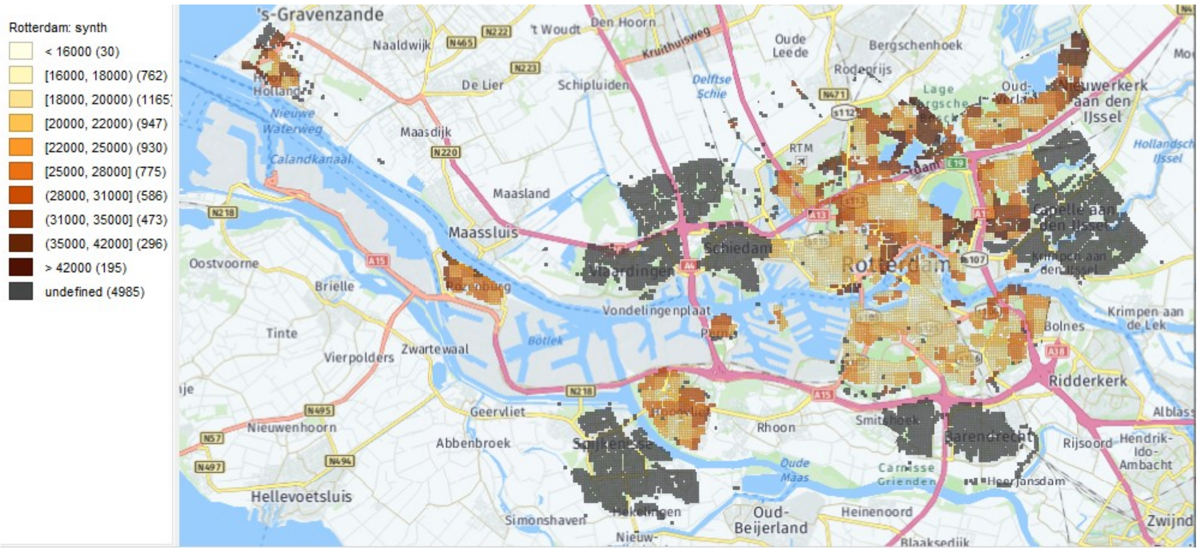


Figure 2.d.29.: Downscaling-result for Rotterdam
Synthetic approach

If suitable covariates are available a regression model can be fitted on level of the larger cells and then be employed to obtain predictions on level of the smaller cells. Applying model selection routines from the R package `olsrr` and employing cross validation to check the predictive performance of candidate models we chose a subset of predictors from the range of covariates from administrative sources listed in Table 2.d.5. A special focus was on model parsimony, as in our problem setting, model parsimony has the additional advantage of creating less empty cells when predicting the median income on the finer resolution level of 100 m grid cells. As a result, we fitted a linear model using HVAL, OWNH and ELEC as covariates. An overview of the fitted model is given in Table 2.d.7.

We obtain predictions on the target level. Results are pictured in Pictures 2.d.30 to 2.d.32. Comparing results to the benchmark approach of synthetic downscaling in Pictures 2.d.27 to 2.d.29, it is obvious that the approach manages to largely retain the broad spatial patterns in the income data, while at the same time introducing some small-scale heteoregenity. It does, however, clearly seem to overestimate smaller incomes. Furthermore, we obtain a large number of NAs due to missing predictors on the target level; overall, for 6956 out of 16011 grid cells no value could be calculated because one ore more of the auxiliary information was NA. In 2101 of these cells, the number of residing people is NA, i.e. these cells are actually empty cell or cells with less than 5 inhabitants. For the remaining 4855 cells, a result would be desirable.

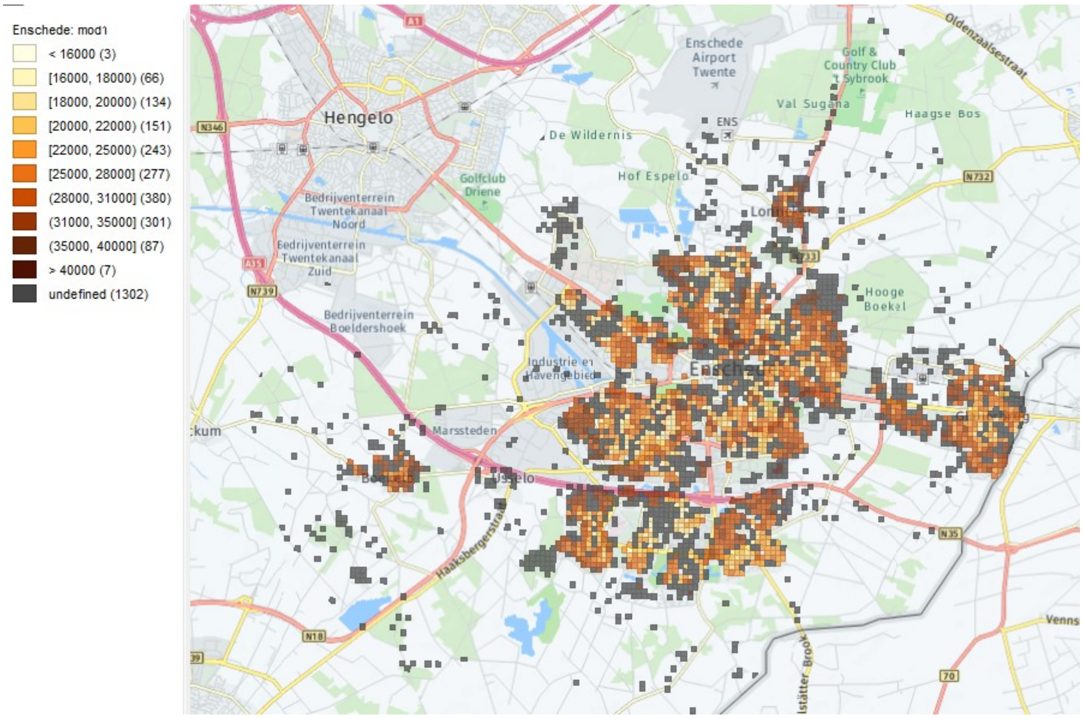


Figure 2.d.30.: Downscaling-result for Enschede
Modell 1 (auxiliary information from administrative sources)

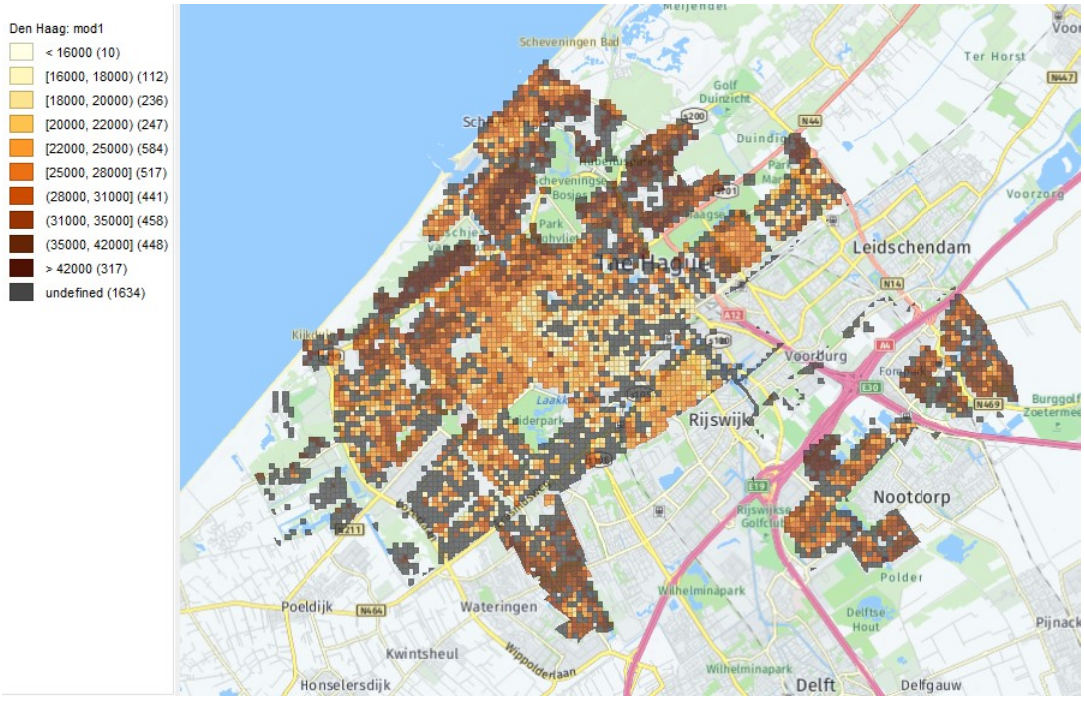


Figure 2.d.31.: Downscaling-result for Den Haag
Modell 1 (auxiliary information from administrative sources)

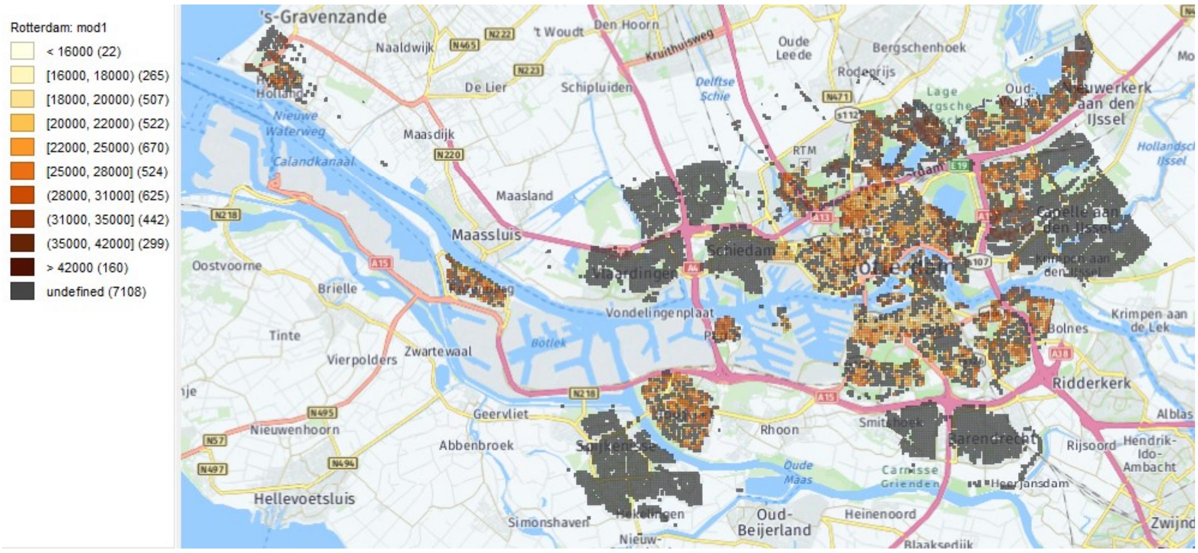


Figure 2.d.32.: Downscaling-result for Rotterdam
Modell 1 (auxiliary information from administrative sources)

We, therefore, now exploit the usability of the remote sensing data indices introduced above. Again, we employ standard model selection techniques and check resulting candidate models for their prediction accuracy using cross validation. Corresponding to the findings of city analyses in Section , we also consider city-specific models as an alternative to one model for all cities. An overview of resulting models is given in Table 2.d.8. It can be taken from this overview that indeed different predictors are chosen for different cities and that overall the coefficient of determination, R^2, can be improved when estimating city-specific models. This is specifically true for Enschede and even more Den Haag. All estimated coefficient are highly significant. While the goodness of fit, overall, is considerably worse than when employing data from administrative sources, we are still able to explain some of the variation in the data.[3]

	All cities	Enschede	Den Haag	Rotterdam
(Intercept)	19532.01^{***}	17063.79^{***}	19523.42^{***}	16583.27^{***}
	(804.45)	(1494.21)	(1172.40)	(1280.03)
BUmean	-13121.75^{***}			
	(1457.08)			
NDVImean		20482.06^{***}	30282.25^{***}	
		(3780.63)	(3622.00)	
NDBImean				-40778.49^{***}
				(5559.51)
R^2	0.08	0.14	0.19	0.11
Adj. R^2	0.08	0.14	0.18	0.11
Num. obs.	925	179	307	439

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2.d.8.: Models using remote sensing data

Downscaling results from this models are presented in Pictures 2.d.33 to 2.d.35.

[3] We also considered the contextualized indices described in 2.d.4.2.2. While the R^2 could be increased for Enschede and Den Haag, estimated coefficients were not significant and prediction accuracy of the models judged by cross validation was worse.

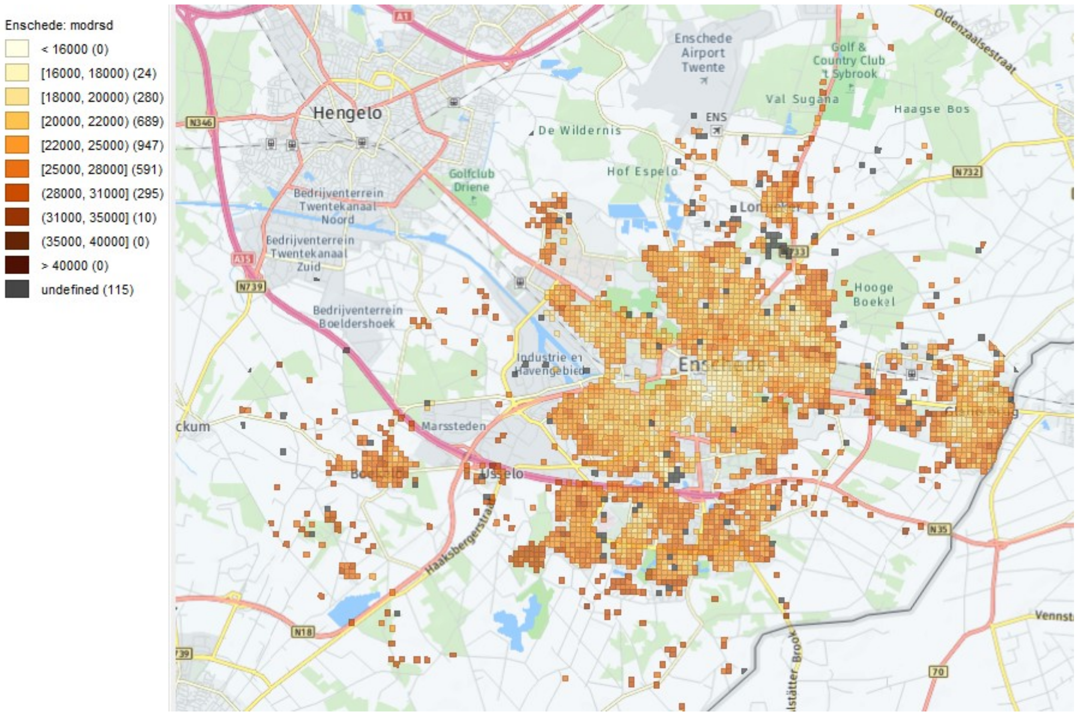


Figure 2.d.33.: Downscaling-result for Enschede
Modell 1 (auxiliary information from remote sensing data)

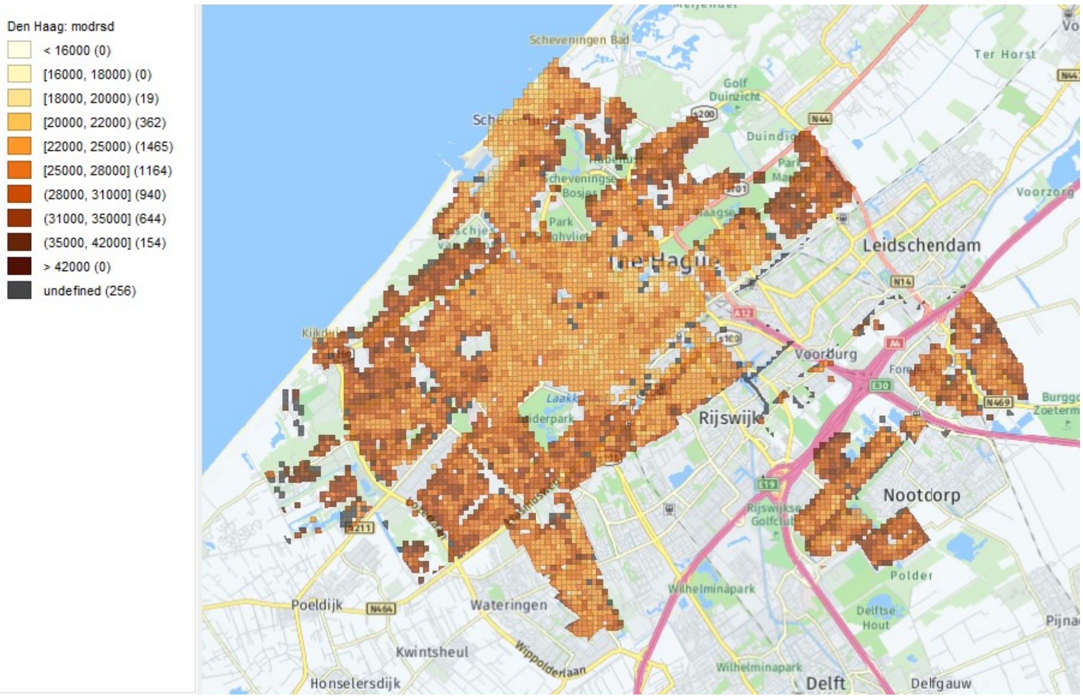


Figure 2.d.34.: Downscaling-result for Den Haag
Modell 1 (auxiliary information from remote sensing data)

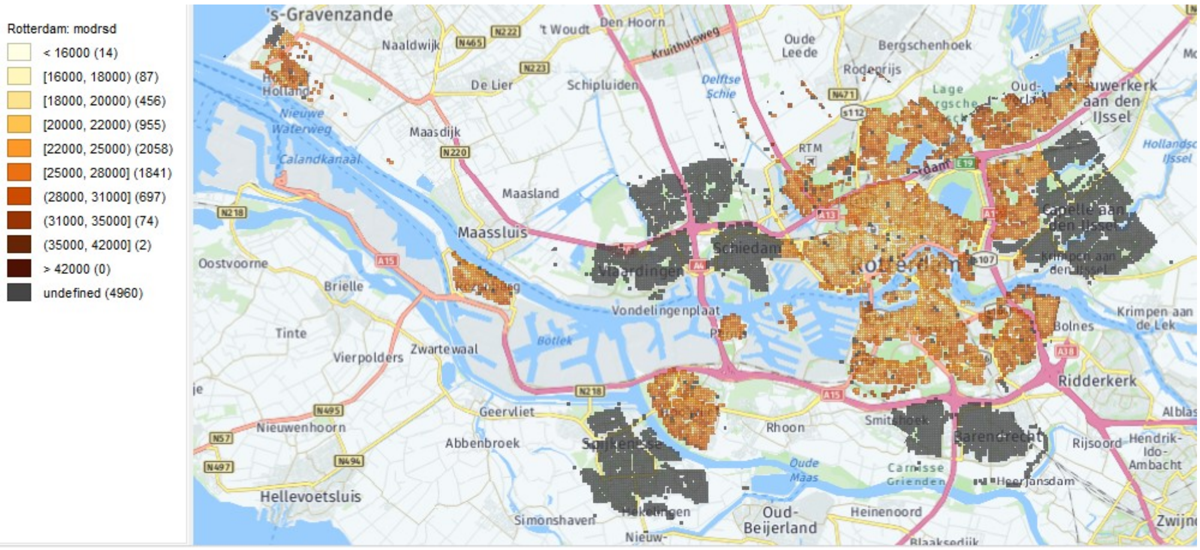


Figure 2.d.35.: Downscaling-result for Rotterdam
Modell 1 (auxiliary information from remote sensing data)

We can take from this images, that overall structures are largely pertained. In Enschede, for example, we detect the patterns of a lower income in the center of the city and areas of higher income in its outer districts. At the same time, the models add the desired small-scale heterogeneity depending on features of the micro-location, i.e the density of vegetation and population. They further produce a result for all non-empty cells in the data set. However, it is also clearly visible that the models, generally, fail to predict the higher and lower incomes and, thus, are not able to pertain the range of the income distribution in the cities under study. This is not an surprising result; while features such as the proximity to parks and the density of construction might determine the quality of living on the micro-level, on level of a whole city other characteristics of the location are of relevance, too. Considering this results, an adapted downscaling approach which uses the larger-scale statistic as a baseline and then adds some small-scale heterogeneity based on statistical relationships of median income and remote sensing indices, i.e. on features of the micro-location, might be a promising approach.

2.d.5. Conclusions

In this chapter, we investigated the opportunities of using remote sensing data in social statistics. More specifically, we analysed the opportunities, prerequisites and limitations of downscaling the median income for the cities of Enschede, Den Haag and Rotterdam using remote sensing data based indicators as auxiliary information. The target level was the very fine resolution level of 100 m grid cells. We present sources of freely available small-scale geo data and related open-source tools for processing the data. Furthermore, we discuss limitations, possible error sources, and respective solutions for these data. Addressing the question of how the information derived from satellite images can be The downscaling is preceded by a careful analysis of spatial patterns in the cities under studies, laying the foundation for the formulation of assumptions about the usability of remote sensing data for modelling the median income.

Finally, the downscaling is performed by first estimating models on level of the larger grid cells and then predicting the median income on the target level from these models. Overall, we can take from this study that on a very low resolution level, alternative data sources can be a promising addition to information provided by Official Statistics. On this low level of analysis, traditional indicators are frequently not available due to confidentiality reasons. Contrary to that, the presented remote sensing data based indicators are available for all cells under study. Furthermore, we could show that, while not having the same explanatory power as information from administrative data, variability in the median income could at least partly be explained by remote sensing based indicators of vegetation and urbanity. We, thus, believe that it is worthwhile to pursue this approach further and to develop adequate methods to exploit this information source.

For the presentation of some results we used maps from the following sources:

Information about the ESRI WoldTopo map see: `https://www.arcgis.com/home/item.html?id=30e5fe3149c34df1ba922e6f5bbf808f`.

Information about the HERE Hybrid maps see: `https://developer.here.com/documentation/map-tile/dev_guide/topics/example-basemap.html`.

2.e. Estimation of the number of people earning minimum wage using a measurement error model

2.e.1. Introduction

Model-based small area estimation (SAE) methods are designed to produce reliable estimates in the presence of very small sample sizes. This is achieved by employing predictive statistical models which combine survey data from different areas with auxiliary information obtained from other sources. Prerequisites for the successful application of model-based SAE are the identification of a suitable model and the availability of auxiliary information which have predictive power for the parameters of interest. Frequently used sources of auxiliary information include registers and census data. However, in many situations where model-based small area estimates are needed, no suitable auxiliary information can be obtained from registers or census data. Hence, it seems natural to incorporate auxiliary information from big data sources. As linking information from big data to individual units from the target population may amount to a tremendous task, it is more common to incorporate big data sources within area level models (cf. Marchetti et al., 2015; Marchetti et al., 2016), where aggregate information on the area level is required. One of the assumptions of the area-level model is that the covariates relate to known population values, which are measured without error and cover the population of interest (Marchetti et al., 2015). Note that these assumptions may be violated by auxiliary information obtained from big data sources, e.g. in the case of using aggregates from social media issues due to self-selection may arise.

Over the last few years, a number of methods have been proposed to incorporate covariates measured with error in area level models. For reviews on these developments see Rao and Molina (2015b, Section 6.4.4) or Pfeffermann (2013). Perhaps the most prominent example is due to Ybarra and Lohr (2008), who derive an optimal estimator when covariates are measured with error under the cross-sectional model using a frequentist approach. Alternatively, Arima et al. (2015) develop a Bayesian approach to deal with covariates subject to measurement errors and Arima et al. (2017) extend their approach to a multivariate model.

In our experimental application, we consider SAE methods to produce regionally disaggregated estimates of the number of people earning the general statutory minimum wage in Germany. While we do not have access to strong predictive covariates from registers, a covariate obtained from another survey is available, motivating us to consider estimation methods accounting for sampling error. Since we want to produce estimates for multiple consecutive years, we additionally want to borrow strength across time using a time series model. An overview on time series models for SAE using area level data is given in Rao and Molina (2015b, Section 4.4.3). While there are many approaches to modeling time series in SAE, we consider a model similar to Datta et al. (2002), where time specific area effects are assumed to follow a random walk. Putting the different components together, we apply a time series model for SAE where the covariates are measured with error. To the best of our knowledge, this issue has not been studied so far in the literature.

The remainder of this part is structured as follows. A description of the surveys from which our data are obtained is given in subsection 2.e.2. Our estimation methods are presented in subsection 2.e.3. The results of the experimental allocation are shown in subsection 2.e.3.1. Finally, subsection 2.e.4 offers a brief summary and discusses our findings.

2.e.2. Background and data description

In 2015, Germany established a general statutory minimum wage. To monitor the impact of the minimum wage the Federal Government has set up a standing Minimum Wage Commission. In the German statistical system, detailed data on the wage structure are provided by the survey on the earnings structure, which is conducted every 4 years. However, at the time the general statutory minimum wage was enacted, the most recent survey data was available for the reporting year 2014, with the next survey scheduled only for reporting year 2018. As there was a need to provide the aforementioned commission with recent meaningful information the Federal Ministry of Labour and Social Affairs commissioned Destatis to conduct a federal statistic according to paragraph 7 of the Federal Statistics Law (Frentzen and Günther, 2017). Under this provision Destatis implemented a survey on earnings for the reporting years 2015 to 2017. A distinctive feature of surveys according to paragraph 7 is that the sampled units are not obliged to participate. Owing to the voluntary participation rather low response rates were achieved, ranging from 6 to 15 per cent (Frentzen and Günther, 2017, Kann, 2018). To produce reliable results despite low response rates a two-step estimation procedure was used, combining an adjustment for non-response with a calibration to known population margins via a generalized regression estimator. This procedure yielded satisfactory results for the total number of people earning the minimum wage in Germany. However, estimates on more granular levels such as for states or branches of economic activity were in general not reliable enough. This motivated an internal research study by Destatis to assess whether better estimates could be obtained by means of model-based SAE methods.

2.e.3. Methods

In our application, the target parameters of interest are the total numbers of people earning the minimum wage at the state level, which we will denote by $\tau_d, d = 1, \ldots, 16$. Since the τ_d are counts, it is common practice to model $\theta_d = \log(\tau_d)$. Thus, our starting point for a single reporting year is the cross-sectional area level model given by (Rao and Molina, 2015b, chapter 6)

$$\widehat{\theta}_d^{CAL} = \theta_d + e_d = X_d'\beta + v_d + e_d, \quad d = 1, \ldots, 16. \tag{2.e.1}$$

Here $\widehat{\theta}_d^{CAL} = \log\big(\widehat{\tau}_d^{CAL}\big)$ denotes the logarithm of the calibrated state level estimates, which is assumed to be linearly related to X_d, a $p \times 1$ vector of known covariates at the area level, β denotes the $p \times 1$ vector of regression coefficients, $v_d \sim N(0, \sigma_v^2)$ denotes the area-specific random effect and $e_d \sim N(0, \psi_d)$ the sampling error, which is generally assumed to be known.

Note that $\psi_d = Var(\widehat{\theta}_d^{CAL}) \approx Var(\widehat{\tau}_d^{CAL})/(\widehat{\tau}_d^{CAL})^2$ by linearization. Under model 2.e.1 the empirical best linear unbiased predictor (EBLUP) for θ_d is then obtained as

$$\widehat{\theta}_d^{EBLUP} = \widehat{\gamma}_d \widehat{\theta}_d^{CAL} + (1 - \widehat{\gamma}_d) X_d' \widehat{\beta}, \qquad (2.e.2)$$

where $\widehat{\beta}$ denotes the best linear unbiased estimator of β and $\widehat{\gamma}_d = \widehat{\sigma}_v^2 / (\widehat{\sigma}_v^2 + \psi_d)$. A predictor for τ_d may be obtained by an appropriate back transformation of $\widehat{\theta}_d^{EBLUP}$.

For our application it is necessary to extend model 2.e.1 in two ways. First, note that we have actually data for three consecutive reporting years, which is why we may want to exploit a temporal correlation to obtain better estimates then under the purely cross-sectional model 2.e.1. Secondly, as auxiliary information we may use the number of people earning less than the minimum wage enacted in 2015 from the survey on the earnings structure for the reporting year 2014. As this covariate stems from a sample survey, we should account for its sampling error as well. We will explore each extension in turn.

To describe the combined model for reporting years 2015 to 2017, it is necessary to augment our notation to include a time index $t, t \in \{1; 2; 3\}$, where $t = 1$ refers to reporting year 2015 and so on. Thus, $\widehat{\tau}_{dt}^{CAL}$ now denotes the calibrated point estimate in state d at time t and $\widehat{\theta}_{dt}^{CAL} = \log\left(\widehat{\tau}_{dt}^{CAL}\right)$. Our extended model now reads:

$$\widehat{\theta}_{dt}^{CAL} = \theta_{dt} + e_{dt} = X_{dt}' \beta_t + v_{dt} + v_d + e_{dt}, \forall d, \forall t. \qquad (2.e.3)$$

In 2.e.3, X_{dt}' denotes the vector of known covariates in state d at time t, β_t is the associated vector of regression coefficients at time t and $e_{dt} \sim N(0, \psi_{dt})$ denotes the known sampling error in state d at time t. Note that unlike the cross-sectional model 2.e.1, the extended model 2.e.3 comprises two random effects. In addition to random effects for the states $v_d \sim N(0, \sigma_v^2)$, random effects for time specific state effects v_{dt} enter the picture. We assume a random walk for the time specific state effects $v_{dt} = v_{d,t-1} + \xi_{dt}$ where the innovations follow a white noise process, $\xi_{dt} \sim N(0, \sigma_\xi^2)$. Owing to the random walk assumption, model 2.e.3 implies persistent deviations from the fixed part of the model, which seems reasonable. Note that model 2.e.3 was originally proposed by Datta et al. (2002) to estimate the median income of four-person family in the states of the USA.

To combine different periods, we may write $\boldsymbol{\theta}_d = (\theta_{d1}, \theta_{d2}, \theta_{d3})'$ as the vector of target parameters under model 2.e.3 from different periods in a given state. The covariates belonging to state are stored in the matrix $\boldsymbol{X}_d = diag(X_{d1}', X_{d2}', X_{d3}')$ and the vector of regression coefficients for all years is given by $\boldsymbol{\beta} = (\beta_1', \beta_2', \beta_3')'$. The covariance matrix of the sampling errors is given by $\boldsymbol{\psi}_d = diag(\psi_{d1}, \psi_{d2}, \psi_{d3})$. Note that this specification corresponds to independent sampling errors across time. The covariance matrix of the random effects follows as $\boldsymbol{\Sigma}_v = \sigma_v^2 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \sigma_\xi^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$. As noted by Bell (2012), the covariance matrix would become singular if either $\sigma_v^2 \to 0$ or $\sigma_\xi^2 \to 0$, and would be ill-conditioned for a large number of time periods, the latter of which is not a major concern in our application with only 3 periods. We can now express the EBLUP under model 2.e.3 as

$$\widehat{\boldsymbol{\theta}}_d^{EBLUP} = \boldsymbol{\Sigma}_v(\boldsymbol{\psi}_d + \boldsymbol{\Sigma}_v)^{-1}\widehat{\boldsymbol{\theta}}_d^{CAL} + \boldsymbol{\psi}_d(\boldsymbol{\psi}_d + \boldsymbol{\Sigma}_v)^{-1}\boldsymbol{X}_d\widehat{\boldsymbol{\beta}}, \quad d = 1, \ldots, 16. \tag{2.e.4}$$

While 2.e.4 accounts for temporal correlation, it still assumes known covariates $\boldsymbol{X}_d$, which is violated in our application.

A conceptually straightforward extension of 2.e.3 is to replace the assumption of fixed covariates $\boldsymbol{X}_d$ by estimated covariates $\widehat{\boldsymbol{X}_d}$. We may want to assume that $\widehat{\boldsymbol{X}_d} \overset{ind}{\sim} N(\boldsymbol{X}_d, \boldsymbol{C}_d)$, which can be justified if the covariates are estimated from a large probability sample. Note that simply replacing $\boldsymbol{X}_d$ by $\widehat{\boldsymbol{X}_d}$ and using 2.e.4 does not lead to valid MSE estimates. For the case of the cross-sectional model 2.e.1, Ybarra and Lohr (2008) show that using 2.e.2 with estimated covariates may in fact lead to estimates which are even worse than the direct estimates. As we are interested in a flexible approach that can be used for different models with a minimal amount of specific adjustments, we decided to adopt a Bayesian framework where we incorporate the uncertainty due to estimated covariates in a similar way to Arima et al. (2015).

In our application, the only readily available covariate information is the estimated number of people earning up to the minimum wage that came into place in 2015 at the state level from the survey on the earnings structure as the only auxiliary information. We denote the logarithm of this quantity by $\widehat{z}_d$. Our final model is as follows:

$$\begin{aligned}
\text{(sampling model)} \quad & \widehat{\boldsymbol{\theta}}_d^{CAL}|\boldsymbol{\theta}_d \overset{ind}{\sim} N(\boldsymbol{\theta}_d, \boldsymbol{\psi}_d) \\
\text{(linking model)} \quad & \theta_{dt} = \beta_{0t} + \beta_1 z_d + v_{dt} + v_d \\
\text{(random effects)} \quad & v_{dt}|v_{d,t-1}, \sigma_\xi^2 \sim N(v_{d,t-1}, \sigma_\xi^2), \quad v_d|\sigma_v^2 \sim N(0, \sigma_v^2) \\
\text{(prior distributions)} \quad & \beta_{0t} \overset{i.i.d.}{\sim} N(0, 100^2),\ \beta_1 \sim N(0, 100^2),\ z_d \sim N(\widehat{z}_d, \sigma_{z,d}^2), \\
& \sigma_\xi \sim U(0, 20),\ \sigma_v \sim U(0, 20)
\end{aligned} \tag{2.e.5}$$

Hence, our model comprises as fixed effects time specific intercepts as well as a time-invariant slope coefficient. Note that our prior specifications for $\beta_{0t}, \beta_1, \sigma_\xi$ and σ_v are non-informative or only very weakly informative. We treat $\widehat{z}_d$ as known and fixed to the corresponding point estimate from the survey on the earnings structure in 2014 and $\sigma_{z,d}^2$ is the corresponding variance estimate and also treated as fixed. We are willing to fix $\widehat{z}_d$ and $\sigma_{z,d}^2$ as they are obtained from a large sample designed to provide reliable state level estimates. Markov Chain Monte Carlo (MCMC) methods are used to estimate posterior distributions of all parameters. The estimates of $\boldsymbol{\theta}_d$ under model 2.e.5 are obtained as the posterior means $\boldsymbol{\theta}_d$ given all other parameters and the direct estimates. Suppose that we have R samples from the posterior distribution and let $\widehat{\theta}_{dt}^{(r)}$ denote the estimate from the r-th sample for state d and time t on the log scale. The Bayes estimate for the number of people earning the minimum wage in state d and time t can be computed as $\widehat{\tau}_{dt}^{HB} = \frac{1}{R}\sum_{r=1}^{R} \exp\left(\widehat{\theta}_{dt}^{(r)}\right)$. Since reliable estimates for the national total number of people earning the minimum wage at time t given by $\widehat{\tau}_t^{CAL} = \sum_{d=1}^{16} \widehat{\tau}_{dt}^{CAL}$ are available, we would like to ensure that our Bayes estimates add up to this number. In general, however, $\sum_{d=1}^{16} \widehat{\tau}_{dt}^{HB} \neq \widehat{\tau}_t^{CAL}$. Hence, we use a multiplicative benchmarking to ensure that state level Bayes estimates add up to the national estimate obtained by calibration.

2.e.3.1. Results

We use JAGS (Plummer, 2003) to estimate model 2.e.5 and to obtain posterior distributions for all parameter estimates. We run 3 parallel chains with overdispersed starting values, were we discard the first 100000 draws as burn-in period. We then obtain 250000 samples from each chain, keeping every 50th sample to reduce the autocorrelation. This leaves us with $R = 3 \cdot 5000 = 15000$ samples from the posterior distribution. We follow standard practices (cf. Carlin and Louis, 2008, Chapter 3.5) to monitor convergence of the MCMC algorithm by inspecting trace and autocorrelation plots for the different parameters as well as the potential scale reduction factor developed by Gelman and Rubin (1992). The inspection of the trace-plots suggests that the parallel chains mixed well. The autocorrelation plots do not indicate the presence of relevant autocorrelation of the parameter estimates. Also, the potential scale reduction factors are very close to 1 for all parameters. Therefore, the convergence diagnostics suggest that the MCMC algorithm converged.

We use posterior predictive checks to assess the validity of model 2.e.5 as described in Rao and Molina (2015b, Chapter 10). To assess the overall fit of the model we used posterior predictive p-values which provide a tool to measure the discrepancy between the posterior predictive values and the observations. In our application the posterior predictive p-values are in the range between 0.23 and 0.56, which does not indicate a lack of fit. Moreover, we also look at the fit of the individual level. Following You and Rao (2002), we compute the share of replications from the posterior predictive distribution which are smaller than the corresponding direct estimates that have been modeled. The mean and median are both very close to 0.5, which does not suggest consistent over- or underestimation in general. Nevertheless, for one state we observe in 2015 consistently higher values using the replications from the posterior predictive distribution. In this case, however, the direct estimate more than triples in the following year which is most likely an artefact due to the sampling error. Interestingly, the posterior mean avoids this implausible jump. Another diagnostic that we use are the standardized residuals from the posterior predictive distribution, which are depicted in Figure 2.e.1. The only residual exceeding an absolute value of 1.5 corresponds the aforementioned unstable state estimate from 2015. All other standardized residuals are small, thereby suggesting adequate fit of the data. In addition to these internal model checks, we also apply an external model check. For this purpose we compute the ratio between the national estimates obtained by calibration and the national estimates that would result by adding up the unadjusted posterior means in a given year. In all three years this ratio is close to 1, which does not lead to major concerns.

Perhaps the most critical assumption in the model 2.e.5 concerns the choice of uniform distributions as prior distributions for σ_v and σ_ξ. Figure 2.e.2 depicts histograms of the posterior

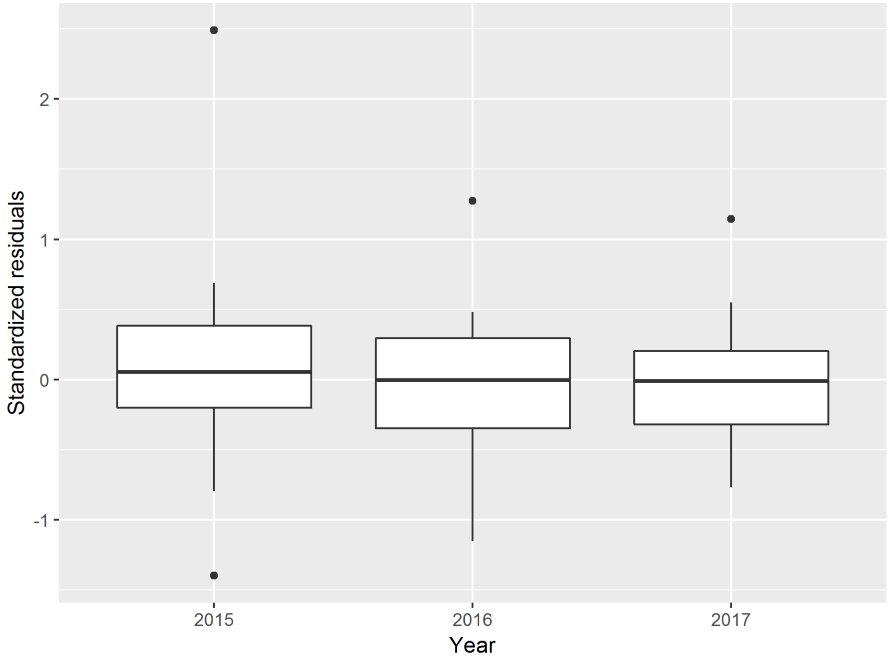


Figure 2.e.1.: Standardized residuals from the predictive posterior distribution

distributions of these standard deviations by chain. It is clearly visible that setting the upper bound to 20 does not effectively constrain the parameter estimates in this application. Note that the posterior means are given by 0.14 and 0.13 respectively. As a robustness check, we consider a specification with half-Cauchy distributions used as prior distributions for σ_v and σ_ξ, which were recommended by Gelman et al. (2006). The posterior means and standard deviations using half-Cauchy priors are virtually identical to those using uniformly distributed priors. Hence, we decide to stick to the earlier assumptions outlined in model 2.e.5.

Our final Bayesian estimates are obtained after benchmarking each year and sample from the posterior distribution to the national estimate for that year. A comparison of the coefficients of variation (CV) under the Bayesian approach and under the calibration estimator is shown in Figure 2.e.3. Here, HB refers to the Bayesian estimates and CAL refers to the calibration estimates. Note that these quantities are conceptually different, as the CV under the Bayesian approach is obtained with respect to repeated realizations from the model 2.e.5, while the CV of the calibration estimates refer to distribution under repeated sampling from a fixed and finite population (Boonstra and van den Brakel, 2019).

The first striking aspect in Figure 2.e.3 is that the CVs using the model-based Bayesian approach are much smaller than those under the design-based calibration approach for all years. Comparing the CVs over time, we note that for both estimation methods the precision decreases over time. The deterioration of the quality of the Bayesian estimates can be attributed to at least two causes. On the one hand, the precision of the calibration estimates which are used as

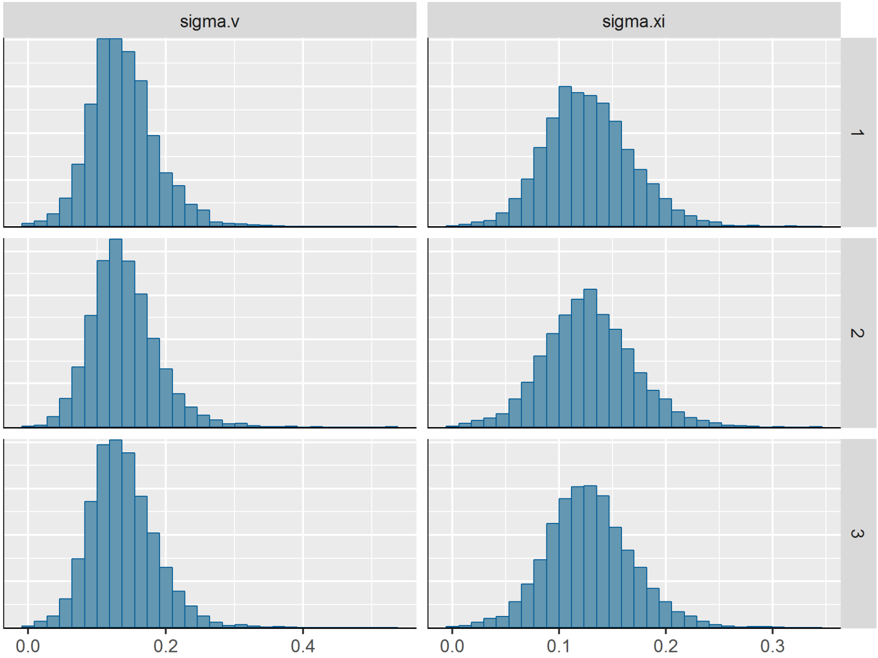


Figure 2.e.2.: Posterior distributions for σ_v and σ_ξ

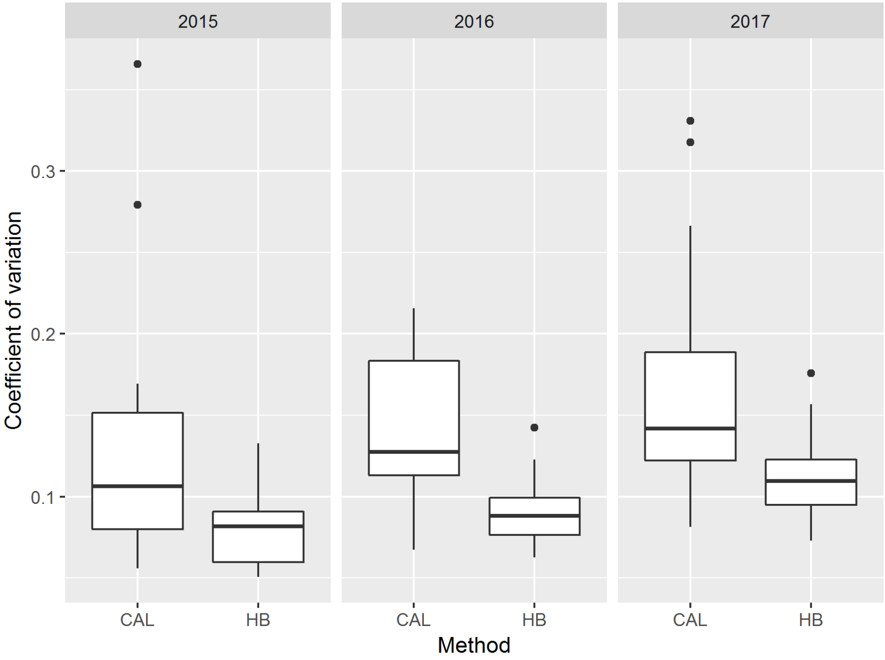


Figure 2.e.3.: Coefficients of variation

an input decreased over time. For the simple cross-sectional model without transformation, the dominating term in the MSE is the shrinkage factor times the sampling variance of the direct estimate. Though our model is more complicated, larger sampling errors may still propagate into larger errors of the model-based estimates. On the other hand, the auxiliary information used in our model refers to the year 2014. While we may expect that it is highly predictive for the year 2015, it seems plausible that the predictive power deteriorates over times as the structures change.

2.e.4. Discussion

We develop a time-series model for SAE where the covariates are measured with error to borrow strength across space and time as well as to account for the uncertainty due to using estimated covariates. In our experimental application, we use this SAE model to estimate number of people earning the minimum wage in German states. In German official statistics, state level estimates from sample surveys are usually obtained using design-based estimation methods. However, due to the low response rates in the voluntary survey on earnings, the available sample sizes at the state level are not sufficient to produce reliable direct estimates. The results of our experimental application are clearly promising, as a substantial reduction of the CVs is achieved using our model-based approach. Moreover, the model-based approach leads to estimates of changes which are much more plausible than those that would have been obtained using the design-based approach.

Note that we consider a Bayesian approach to account for the uncertainty due to estimated covariate information. While it might have been possible to stay in the frequentist paradigm and extend the methodology by Ybarra and Lohr (2008) also for time-series models, we did not pursue this approach further. Our main reason to prefer a Bayesian approach is that we are interested in a generic approach which is applicable to different model specifications without the need for major specific adjustments. In addition to our final model specification 2.e.5, we explored a number of alternative specifications to find a suitable model that produces good estimates which are robust under modified assumptions. Changing the prior distributions for the variance parameters from uniform distributions to half-Cauchy distributions had virtually no effect on the parameter estimates. As another alternative, we considered a specification where the sampling and linking model are unmatched, i.e. where they cannot be combined into a linear mixed model (You and Rao, 2002). Interestingly, the posterior means before benchmarking were generally a bit smaller using the unmatched sampling and linking model in comparison to the matched case, consequently requiring a stronger benchmarking adjustment. Finally, we also ran a model specification where we pretend that the auxiliary information were measured without error. The point estimates assuming known auxiliary information were very similar to those accounting for the uncertainty. A possible explanation for this finding is that the auxiliary information are obtained from a large survey where participation is mandatory. Therefore, the sampling errors in the auxiliary information are much smaller than those of the calibrated estimates from the voluntary survey.

Moreover, we would like to note that the German system of surveys on earnings is currently being revised. Under the new system which is supposed to be used for the first time in the reporting year 2021, there will be a single mandatory survey where information from the selected establishments will be collected on a monthly basis. This new survey will provide both short-term statistics as well as structural statistics on earnings. Hence, calibrated estimates of the number of people earning the minimum wage at the state level can be expected to be much more precise in the future compared to this application.

In our application, the measurement error of the covariates is due to using sample estimates from a larger survey in place of register information, which is scarce. However, measurement error models may be useful as well if covariates were obtained from big data sources, where standard assumptions that these data sources fully cover the relevant target population and are measured without error are unlikely to hold. An example in this regard, the truck toll mileage index produced by the Federal Office for Goods Transport and Destatis is strongly related to the industrial production in Germany. As the truck toll mileage index is available a few days after the reporting period it can be used as an early indication of the short-term economic trends (Destatis, 2020b). Another potential application is to employ mobile phone data as an indicator for regional mobility behavior in light of the current Covid-19 pandemic (Destatis, 2020a).

3

Further research needs and best practice guidelines

3.a. Research needs

3.a.1. Research needs for advanced applications using new data sources

This deliverable addressed the use of satellite data for the estimation of well-being indicators on regional levels. In the course of the MAKSWELL project, it could be shown that considering such non-standard data sources marks a powerful addition to quantitative socio-economic research. Given the unprecedented increase in data volume and the rapid advance of statistical methodology over the past decades, the findings of this project suggest a very broad scope for future research. In what follows, we point out several current research fields in statistics where the advances of the MAKSWELL project could be included to enhance the methodology for empirical studies based on non-standard data.

The first field is multi-source estimation. Researchers might often need to choose between a variety of surveys that are relevant for their research question. In this context, different techniques to combine the information of multiple surveys exist, as for instance proposed by Tighe et al. (2010), Rao et al. (2008), as well as Schenker and Raghunathan (2007). Against this background, Roberts and Binder (2009) distinguish between two general approaches: a separate approach, where an overall estimate is obtained by combining the separate estimates based on the single surveys, and a pooled approach, where the individual records from all surveys are combined in order to perform the estimation on the pooled sample. With regards to the MAKSWELL project, the ideas of multi-source estimation could also be extended to non-standard data sources. In particular, satellite images and remote sensing data could be used to augment basic survey records for socio-economic analysis. With this combination, the researcher would have insights into both person-related and environmental aspects, which is likely to improve statistical modeling for context-related analysis.

A second closely related field is meta-analysis. The term meta-analysis describes statistical techniques to combine related study results by synthesizing summary statistics, such as effect sizes and standardized mean differences, correlation coefficients, or odds ratios. Following Roberts and Binder (2009), this classical conception of meta-analysis (so-called aggregated data meta-analysis) can be understood as an alternative approach for combining survey information. The advantage of this perspective is that it allows to efficiently combine insights from large numbers of observations that would otherwise push conventional data processing software to its limits (Silva-Fernández and Carmona, 2019). Depending on the chosen methodology, the computational burden for statistical analysis grows exponentially with the number of considered data records. In practice, especially when dealing with big data sources, this quickly leads to infeasible calculations. Meta-analysis provides a natural solution to this problem by combining summary statistics rather than actual observations. With regards to the MAKSWELL project, this is particularly relevant when dealing with satellite images. Selected approaches of meta-analysis could be applied to (i) draw relevant information from multiple data subsets and (ii) to pool their insights for an efficient data analysis from different sources.

This leads us to the final field, which is big data analysis. Statistical methods must increasingly meet the challenge of analyzing and using big data records. The “four V” definition is nowadays widely accepted to briefly summarize the characteristics of big data, that is, volume, variety, velocity and value. For an overview of big data, see for instance Chen et al. (2014). In the future, official statistics will also have to deal more and more with how the use of big data can be used to improve analyses. Daas et al. (2015) give an overview of the opportunities and challenges official statistics face in light of big data. They also give two case studies on the use of big data in official statistics. In the first case study they show the use of traffic loop detection data for illustrating traffic movements in the Netherlands. In the second case study they study social media messages as they give interesting information on persons opinions towards certain topics. The use of big data for official statistics is closely linked to the fields of multi-source estimation, non-probability sampling and the analysis of missing data.

This is why several new powerful methods that are nowadays commonly used in this field may also be applied on satellite data for socio-economic research, as the methodological issues and primary data characteristics are similar. In particular, regularized regression analysis marks a relevant addition to the methodological landscape on that regard. Regularized regression methods, such as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) or the elastic net (Zou and Hastie, 2005), are techniques that can be efficiently applied on large data sets. The main idea of these methods is to extend the loss function that is minimized for model parameter estimation by a penalty term that contains the model parameters to be estimated. Depending on the chosen penalty, minimizing the extended loss function yields a sparse solution to the minimization problem. This implies that model parameters corresponding to covariates that are irrelevant for the target variable are automatically set to zero in the estimation process. That is to say, regularized regression performs an automatic variable selection, which is useful when dealing with high-dimensional data sets, since they typically do not allow for standard exhaustive variable selection algorithms. Furthermore, the numerical procedures that are typically applied to solve regularized regression problems do not require matrix calculations, which reduces the computational burden drastically relative to standard methods.

In addition to their computational advantages, regularized regression has also been found to be robust against measurement errors and noisy data (Bertsimas and Copenhaver, 2018a). Uncertain data records are a common problem in big data applications. With regards to satellite data, the problem is particularly relevant, as the images are often affected by atmospheric disturbances, clouds, and other weather phenomena. Using regularized regression on these data sets allows for a robust analysis of the contained records and could then yield reliable insights for the socio-economic research question at hand.

3.a.2. Research needs for regional price indices

The design of the UK's CPI is quite typical amongst developed economies, in that it has some probability stages combined with some purposive sampling. It is clear that overall sample sizes for baskets/weights and price quotes are not sufficient to support direct calculation of regional indices, and the use of small area estimation seems to offer only a small, though important, gain. Therefore if regional temporal consumer price indices are an important component of assessing wellbeing, it will be necessary to invest in further data collection. For prices in some commodity groups, this information may be obtained directly as a result of current efforts to obtain and use scanner data in the construction of price indices. A similar administrative data approach may also be possible for weights (and indeed Statistics Norway has already abandoned the household budget survey as a source of basket and weight information, judging that its low response rate and measurement errors make it less satisfactory than alternative, administrative data sources).

3.a.2.1. Future research: regional indices

The discussion of the conceptual framework for regional consumer price indices in Chapter 2.a.2 identified a number of areas where our knowledge is incomplete and some development of new methods or experimentation in the application of existing methods would be beneficial. These are:

1. Spatiotemporal index: Section 2.a.2.1.2 distinguished between regional spatial price indices such as PPPs and regional temporal price indices designed to measure inflation in different regions. It is possible to combine these in a spatiotemporal price index which would allow simultaneous assessment of changes by region and across time. In order for comparisons to be transitive, it would need to be produced using a suitable methodology such as GEKS (the Èltetö– Köves–Szulc method (Èltetö and Köves, 1964, Szulc, 1964a, ONS, 2011)). This method has been applied for regions at a single time in PPPs, and has also been proposed for multiple times for a single index. But in a regional temporal index it would have to be applied to region × time, giving many comparisons. Further, each time a new period was added to the series, the calculation would need to be redone, and this would result in revisions to all the preceding periods. However, there do not seem to be any attempts to calculate such an index reported in the literature, so it would be useful to follow this process and investigate the impact of the revisions. They might be expected to become small for times far in the past, but an unusual change in the current period might have an unexpected effect. An implementation of this procedure is therefore recommended for future research. There are some examples of accounting for spatial and temporal variation in hedonic models for price indices (see Nappi-Choulet Pr and Maury (2009) and references therein), but these seem to be of limited use in a CPI.

2. Accuracy of compositions/baskets. In Sections 2.a.2.2 and 2.a.4.2.2 we consider the estimation of the regional baskets based on survey data. But how should we assess the accuracy of estimating a complete composition (basket), in this case of consumer pur-

chases, a set of proportions which must sum to 1? There are related ideas in the accuracy of the national accounts which could be applied, and there are methods for regression analysis using compositions, but further research into an approach for quantifying the accuracy of a composition would be valuable.

3. Small are estimation of baskets: Section 2.a.4.4 uses smoothing and small area approaches to estimate the weights for a regional price index, which implicitly captures many of the properties of a regional basket of goods and services. Small area estimation uses a model to borrow strength across different areas, so as to reduce the variance, but also introducing some bias. The relative importance of the original data and the model depend on the sample size of the regular data. The effect is to 'shrink' the direct estimator towards the mean for the model. This sort of approach has been applied variable by variable for each product in Section 2.a.4.4, but some normalisation is needed to return to a composition, and it is not clear that the resulting composition would be shrunk towards the mean composition. Instead there are versions of GLMs that deal with compositions (the standard approach (Aitchison, 1986) based on a transformation, but also a new approach (Sammut, 2016) fitting the compositional data directly with a suitable error function). It would be interesting to investigate these models for small area estimation of baskets).

   An alternative small area approach is to use SPREE type estimators and their generalisations (Luna et al., 2015) to estimate the region × expenditure table based on the latest LCF and some older, more accurate information (maybe constructed from a longer run of LCF data).

4. Small area estimation of weights: this is almost the same as small area estimation of baskets. The estimated expenditure on commodity groups, particularly at detailed levels of the COICOP classification, are likely to be very volatile when based on LCF responses. One possibility is to pool multiple years of LCF data to increase the sample size. Another is to consider a small area estimation approach for the weights. Ideally this would be constrained to the overall expenditure total, and this could use some benchmarking or a SPREE type approach.

5. Consistency of national and regional indices: as set out in chapter 2.a.2 the regional CPIs are not consistently constructed (they have different baskets and weights), and therefore cannot be aggregated to any sensible higher level index. However, taking the national CPI as a benchmark could potentially allow consistent system of regional and national indices to be constructed. Further thinking on the properties and desirability of such a procedure would be interesting.

3.a.2.2. Future research: quality measures

It is clear from chapters 2.a.5 and 2.a.6 that producing quality measures for price indices is a major challenge, and despite the considerable efforts that have gone into developing the methodologies, there are relatively few implementations, and even fewer of those which result in regular estimates of quality measures.

There is a considerable challenge in defining what processes are leading to the sampling variance in a price index, where the ideal target is not unambiguously defined. However, an indicator of accuracy, taken using a pseudorandomisation approach as used here (and in many other examples), is an important statistic for users, as identified by McCarthy (1961) and more recently for the UK by UK Statistics Authority (2016). In this respect it is surprising that economists have focussed almost exclusively on measurement errors, and not sought to understand the sampling variability. The provision of more detailed information on variances should stimulate a discussion about the different elements of the quality of a CPI, and what level of accuracy is required in a statistic which has so many uses, and which is clearly visible to the casual user because it affects their pay, and the measures the prices they pay in the shops.

The model-based approach to variance estimation does however have some resonance, because of the difficulty in defining what a design-based sampling variance means for a CPI. A comparison of design- and model-based procedures for the same index would be a very useful piece of further research.

The construction of regional indices is definitely easiest when it is a by-product of the national calculation. This means having a regional element of aggregation which would require some redesign in many countries. Where this is not possible, then the variance estimation methods can be applied to the regional indices. If these require extra modelling (for example through small area estimation), then it may be a considerable challenge to produce effect estimators. Then the model-based approach may be an easier option (especially if research reveals it to measure a similar property to the design-based variance calculations.

There is considerable need for research on the methods for variances of CPIs, and particularly for more experience with their application in a range of different index designs.

3.a.3. Lessons from remote sensing applications

Satellite data are not going to be the solution of every problem. We believe that in the greater scheme of social statistics they will find places of importance in applications, but not every problem can be solved and not every dataset be replaced with such data. The study of remote sensing data for estimation of local level poverty showed opportunities, but also problems and limitations of such approaches and data. The application was a pilot study and leaves many opportunities for expansion. There are, however, already some lessons to learn from this application. Generally, we conclude that remote sensing and satellite information can help also in social statistics. How and to what degree is again determined extremely by the exact research situation.

However, much research is going to be needed to better understand which data to use and how to best use such information, specifically for such complex figures as poverty indicators. In this section we formulate some lessons we gathered from this application, specifically.

•Meaningful relations:

It is essential that remote sensing information are found which are meaningfully relatable to social data of interest also theoretically. This is not a simple task considering the sheer number of candidate data and possible combinations. Without this background no lectures can be drawn and no recommendations be formulated for political parties. This relation will commonly be a very indirect relation and we do not think that remote sensing applications are going to get near the quality of survey based data in any near future. However, many other ways might be found to assist statistical analysis which are currently unexplored.

•Modelling:

Another hurdle are the methodological questions remaining open. This might root in the difference of the underlying philosophy of remote sensing and social statistics. Remote sensing strives to capture and record the individuality of every place. In statistical models however, we seek to identify more simple, generalisable relations which can be applied in other places or times. In what way these perspectives can be brought together is expected to bring forth interesting new approaches.

This idea has been expressed for example by Davison et al. (2013) who believe that the Gaussian, mean focused approach might be ill fit for geographical problems specifically rare events. If this is true, the general system of hypothesis testing based on the central limit theorem would have to be re-evaluated. The application by the CBS and Trier University shows indication that satellite indicators should at least be treated differently then other classically collected data from CBS from a modelling point of view

The satellite indicators express information about the corresponding pixel and do not consider general relations. While the expression of spatial relationships using a Spatial Durbin Error Model improved the model using satellite data greatly, it did not change the model fit when using official statistics data. Data collected by surveys are already the result of the greater system of relations and account for many spatial aspects. The price of housing accounts for the attraction of the neighbourhood. Here we believe that using satellite indicators in models will commonly require an explicit spatial component, while these might not be necessary in other data.

•Uniqueness of areas:

Very general relations, such as in Gosh et al. (2010), between night time lights and GDP on a global scale might hold to some degree. But the more local we investigate structures the more unique they become. Such we believe that it will be difficult to use relations to satellite information from a study area with plenty of data to other areas with lack of data. Even within the Netherlands, the relation between median income and satellite indicators were widely different between the cities. We believe that most satellite applications need area specific fine tuning. Such an estimation of median income in a different city can only be successful if many implicit information are accounted for in the modelling progress. The question of applicability of models fitted in one area to another area remains open for upcoming research.

While the differences in the local level might be differentiable, the entirety of history which made a city district interesting and expensive to live in does not show in satellite data. Two possible solution came to our mind for further research:

- Mixed effect modelling might be the easiest methodological adjustment. Assuming that compositions of districts do not change quickly (although do shift) a fixed effect model will allow to define a base level and the estimation would centre around the local and time differences. This does require panel information. For satellite data this exists naturally, for official statistics data this can be difficult.

- Spatial hierarchical composite estimation: When we try to handle nested area problems such as in the application by UT and CBS, a composite estimator might be able to hierarchically align the grid cells by a more global measure of level difference and a local component differentiating smaller differences within greater area of interest.

•Satellite Data quality:

Our application focused on freely available data only. A comparison to other data sources would however be interesting. We believe that the relation of satellite indicator and target variable is more important then the spatial granularity of the used satellite data. To what degree finer spatial resolution is helpful for an else wise similar approach would be an interesting further research question. It would also guide research on whether investment in non-open, high resolution satellite data might be worth while.

•No islands:

When we model variables with satellite data, it will often be possible to think in greater pictures. Any city and house is part of an greater network. The availability of satellite data over greater areas is one of the advantages of these data. Spatial models if they are needed for satellite applications do however have a specific problem with geographic islands (see (LeSage and Pace, 2014)). Common spatial models are not designed to handle so called islands - areas without neighbours- in the spatial weights matrices. This also prohibits to investigate city overarching applications if no common spatial matrix can be constructed. With remote sensing information it is usually no problem to include further areas, even when interested in just a specific city.

Unfortunately this are not always met by the official statistics data. If we want to explore the suitability of satellite data we need geocoded data of a great area from official statistics. The problem occurs in the combination of missing data and spatial modelling attempts. For extensive spatial model research it would be interesting to use national wide data to connect all the areas. Confidentiality reasons will however often prohibit such application in more rural areas.

3.a.4. Handling measurement errors for the estimation of regional indicators

For the estimation of regional poverty and well being indicators reliable data sources are of utmost importance. Unfortunately, surveys are in general designed to allow for precise nation-wide estimates only. On lower regional level, the sample sizes are typically very small and thus direct estimators like the Horvitz and Thompson (1952) estimator and the GREG (Särndal et al., 2003) have high variances. Even though they are unbiased, or at least asymptotically unbiased, in the small sample case, most realizations of the sampling design will lead to unacceptably large estimation errors. For improving the estimation of regional indicators in this context, small area estimation methods have been proposed. For an overview on small area estimation the book by Rao and Molina (2015b), the review articles by Pfefferman (2002), Pfeffermann (2013), and the discussed paper by Ghosh (2020) are highly recommended. Small area estimation methods are mostly model based and, hence, make use of a statistical model to stabilize the regional indicator estimation.

For this purpose, they need reliable covariates that are correlated with the indicator of interest to gain predictive power of the statistical model. As unit-level information on the whole population of interest is typically difficult to obtain in socio-economic applications, we will focus here on area-level small area estimators. These estimators only need data as input, that is aggregated on the regional level of interest. Aggregated covariates are normally obtained either from registers, a census, big data sources, other surveys or even the same survey of which the indicator of interest is measured. One basic assumption in these area-level models is, that the covariates are measured without error. This is in contrast to the indicator of interest, which is assumed to be estimated with an unbiased estimator and that the variance of the estimates is known but so large, that the estimate is not reliable. When obtaining the covariates from surveys, it is clear that the measurement error free assumption is not valid any more. In the following it is discussed, which kind of errors may arise when producing covariates for area level small area estimation methods.

When considering measurement errors in the covariates two main kind of errors can be assumed. First of all, one can assume, that the errors follow a certain known distribution. When estimating the covariates from surveys using classical estimation methods, then at least the asymptotic distribution of the estimates is known and can be estimated. For example, the asymptotic distribution of the HT is normal (Berger, 1998), hence assuming a multivariate normal error distribution in the estimated covariates is plausible (Wood, 2008). By accounting for this it is often possible to derive analytical estimators that account for the uncertainty induced by the covariates with measurement errors. Some publications make use of this idea and also provide mse estimators (Burgard et al., 2019, 2020).

In the last years the use of registers and big data sources like web scraping, online surveys without underlying probabilistic framework and satellite images become more and more popular as sources for covariates in statistical modelling. However in these cases, the measurement errors may typically not be assumed to be simply normally distributed. Registers depend on the persons that fill the register, and every person may fill it slightly different. Also definitions may change along different countries, whilst having the same variable names in the register. Further, over- and undercounts are long known error sources in registers (see, e.g., Burgard and Münnich, 2012). All these errors, are difficult to explain by an theoretic distribution, and hence analytical derivation as in the case above do not apply. A knew strain of measurement error treatment was developed by Bertsimas and Copenhaver (2018b), Burgard et al. (2019). In this framework, errors in the covariates are treated to be in a certain, preliminarily unknown range. By making use of a connection between robust optimization and regularization, the range of plausible errors can be obtained by cross validation, again allowing analytical derivation and plausible upper bounds for the mean squared error of the estimates (Burgard et al., 2019).

Burgard et al. (2019) propose a novel approach to account for the selectivity of health records. This approach could, given the necessary additional data sources are available, also be expanded to big data sources and therefore needs further consideration in this context.

3.b. Recommendations for future use

Local poverty measurement can be considered as a prototype for modern indicator methodology. In the following we summarize the lessons learnt and prospect some guidelines for best practices implementation for transferring methodology.

Best practice recommendations presented here follow a strategic vision that aims at:

- enhancing statistical information through the full integration of administrative sources, survey data and new sources (BIG DATA)

- pushing statistical research toward methods which allows for the measurement of the accuracy of the results when integrating traditional and new Big data sources, as scanner data and satellite imagery.

This is in line with the scenario outlined in recent years in Official Statistics to provide lively and updated indicators for policy action. This on one hand aims at the integration of administrative and survey sources within thematic statistical registers, on the other hand it aims to enrich traditional sources with new emerging data, BIG DATA. For local poverty measures mainly scanner data and satellite imagery have been used.

The goal under the Makswell project is to enable a monitoring and a vision of the wellbeing and living conditions of the population, the individuation of vulnerable groups and the focusing on the more critical areas for poverty. The purpose of the usage of the new data sources is to develop an agile and at the same time rigorous response capacity by combining the use of traditional sources with the treatment of new sources/Big Data, these due to their information potential and characteristics can guarantee this goal.

The three guidelines, emerging from our research on which to intervene to enhance the use of new BIG DATA sources in this direction, concern methodological, technological and information scouting aspects.

3.b.1. Methodological aspects

Concerning the examples described here for Local poverty measurement we will separate the methodological aspects into two categories: one referred to (i) statistical methods used in the analysis, the other to the (ii) methods used in the data production process. As it is essential to ensure continuity and adequate levels of quality of the statistical outputs, it is necessary to invest in the definition of new methods or in the amending of existing methods, evaluating the possibility of using the new alternative sources/BIG Data to integrate/replace traditional sources.

i Statistical methods:

In the local estimation of poverty indicators the approach to inference has been mainly the model based, model assisted approach, even if the more traditional design-based approach has been used. Small area estimation techniques have been used in the definition of local expenditure to weight prices in the CPI and of local poverty rates. Also in the application using satellite data small area model have been used. The accuracy of the CPI requires further development of the model-based approach as also the synthesis classification of error sources shows. Most national CPIs incorporate probability sampling in some stages of selection of prices and in the calculation of the weights, and the replication based approaches provide a natural way to estimate this variation and the variation due to non-probability but replicable parts of the procedures, but research on nonprobability sampling and the measure of the potential bias is still open. Quality measures of the results especially when obtained integrating survey data and new data sources are difficult and requires coordination between traditional and new data sources.

ii Methods used in the production of the data

As it concerns the methods used in the production of the data used, we note that the existing survey data, administrative archives and Big data sources are still separate and their usage requires a big effort in standardization and research on their integration.

- Scanner data are an innovative data source for local and national price indices and the estimation of local price indices, but they need to be integrated with data from surveys on prices and administrative archives. They provide several advantages that derive from the detailed information available about sales and quantities at weekly frequency, GTIN by GTIN, outlet by outlet throughout the entire national territory, but their usage for producing Spatial Price Indexes requires a revision of the data collection in HES.

- Promoting the development of ”fact checking tools” is important in order to determine data veracity and correctness; as an example in the two-weeks Diary (aimed at investigating the more frequent expenditures referred to large consumption products) of Italian Households Budget Survey, HES, since 2015 Istat asks households to indicate, in a dedicated section, the type of outlets where they have purchased a list of 25 products. This is relevant to determine the coverage by scanner data of the markets of the poor.

- A new conceptual model must be developed to integrate with existing architectural schemes, with the development and implementation of a new reference architecture (including the definition of metadata and privacy preserving techniques) and the adoption of new data and process governance models, including ways to use confidential data from alternative sources and to redistribute centrally accessed data.

- In addition, machine learning solutions conceived by teams with experts in statistics must be defined to enable the use of data collected through smart devices, as mobile phones and also satellite imagery and barcode reader which otherwise cannot be analyzed with traditional methods for quantity and speed.

3.b.2. Technological aspects

Another fundamental area of interest for NSIs on which the intervention is needed concerns the IT platforms that allow the processing and enhancement of new data sources (BIG DATA). These IT platforms, developed in a strictly integrated way with new methodologies, have the task of ensuring the right robustness in the integration and analysis of large amounts of unstructured data. In this context, cooperation at European level among National Statistical Institutes and Universities for the evaluation of already developed tools or the joint creation of new solutions is crucial, to provide the necessary capabilities, operational experience, methodological and regulatory guidelines, and testing infrastructures required for a stepwise integration of big data sources into the production of official statistics. In agreement with the vision of the ESS.VIP on Big Data project launched by ESSC, projects should foresee:

1. Short-term actions aimed at building capacity to harness big data sources and delivering first results on the use of big data as an auxiliary source for the production of official statistics. This cannot avoid the cooperation with the University and with experts in survey methods to define useful metadata and to measure the uncertainty and then the accuracy of the results.

2. Medium-term actions to create the legal, technical and statistical infrastructure for systematic use of big data sources in different domains of official statistics;

3. Long term vision to favor full integration of big data sources into the regular production of statistics and the statistical information architecture in a multisource framework.

It should be highlighted that, in this context, European initiatives of great importance are underway, for example the ”Trusted Smart Statistics Center” is being designed, which will collect various BIG DATA sources at European level in the context of the topic ”on line job vacancy”. This new ’instance’ of data collection could collect data at a European level and then allow consultation and analysis by developed tools for both individual states and citizens.

3.b.3. Scouting of new data sources and of new statistical methods

A third important area concerns the scouting of new data sources, and of new statistical methods to obtain more effective and local poverty indicators.

Our main recommendations follow:

- Multi-source estimation has to be extended to non-standard data sources. In particular, satellite images and remote sensing data could be used to augment basic survey records for socio-economic analysis. With this combination, the researcher would have insights into both person-related and environmental aspects, which is likely to improve statistical modeling for context-related analysis, under model based approach.

- Small area estimation techniques seem crucial in building price indexes at local level (estimates of local expenditure to weight prices) but research is needed on the accuracy of the resulting price indexes.

- Probability and nonprobability sampling techniques are mixed in data collection on prices. Making inference using probability and nonprobability sampling is an emerging issue for additional research, to define a statistical data production process on prices integrating the scanner data source and measuring the accuracy of the results.

- Meta-analysis in the analysis of satellite imagery. The term meta-analysis describes statistical techniques to combine related study results by synthesizing summary statistics, such as effect sizes and standardized mean differences, correlation coefficients, or odds ratios. In practice, especially when dealing with big data sources, this quickly leads to infeasible calculations. Selected approaches of meta-analysis could be applied to (i) draw relevant information from multiple data subsets and (ii) to pool their insights for an efficient data analysis from different sources.

- Exploitation of existing data sources (integration of administrative archives) and dialogue between structured and new unstructured data sources. Data collection by surveys must be coordinated with the usage of scanner data sources with more attention to data from IOT (Internet of Things) devices and smart devices for environmental and behavioral analyzes, to ”Mobile Network Operator data” for human presence and mobility statistics and satellite imagery. In this it is essential to design ”fact checking tools” in order to determine data veracity and correctness and dialogue with currently collected survey data (HES).

To explore the areas described, we recommend the development of projects at national and European level that involve the Statistical Institutes, the Universities, the Public Administration and the Private sector in a logic of cross contamination in which each actor, with his own capabilities, contributes to the definition of new integration and analysis methodologies. From this point of view, it is emphasized that numerous projects are underway at European level in the field of ESS Trusted Smart Statistics (shortly TSS), in particular the domains on which priority are given are: HBS (Household Budget Survey) and HETUS (Harmonized European Time Use Surveys). The orientation of the surveys towards smart surveys needs the cooperation with the University to define the quality of the collected data (ES: at Istat level, the surveys that could be oriented towards smart surveys concern: ICT, Labor Force, Trips and Holidays Survey). In this context, we envision that developing plans for creating agreements for partnerships with private data holders, possibly based on privacy-preserving computing technologies and advanced data disclosure control. It is expected that this collaboration will maximize the relevance of trusted smart surveys and the quality of outcome in their respective domains, with a view to ultimately enhance the final statistical output.

This kind of effective collaboration strategy support the importance of Trusted Smart Statistics in order to have well-informed citizens in a data-driven society and the need to use new alternative data sources. This includes raising the awareness among National Statistical Institutes of the importance of their commitment and among decisions makers concerning the use of third-party data for the public interest.

The study of Local poverty measurement raise also the issue of privacy protection of the data used and obtained by scanner data from Retail Trade Chains, and by remote sensing. For this reason, it is recommended to intervene in the creation of flexible services that include the management of data confidentiality ”on the fly”. It is also necessary to develop a targeted communication that clearly introduces the new methodological approach to be adopted and the opportunities that the new sources offer also developing a strategy to promote participation by citizens, through a coherent combination of public communication and individualized incentives.

It is essential to invest in:

- Quality and metadata frameworks aimed to provide quality assurance and data standardization to favor a wider integration and use of multiple data sources (multi-source statistics and multipurpose sources) in the production of official statistics and wellbeing and poverty measures;

- Promoting the development of ”fact checking tools” in order to determine data veracity and correctness; including this in the process of construction of the measures of accuracy of the estimates of the indicators.

- Privacy preserving techniques: technical and methodological solutions for privacy-by-design approaches, such as Secure Multiparty Computation or other cryptographic methods enhanced by advanced automated Statistical Disclosure Control features;

- Artificial intelligence/machine learning/automated solutions for dealing with metadata with more attention to ethical aspects, replicability and transparency of both the algorithms used and the training techniques;

- Promoting the use of the Application Programming Interface (API) paradigm, by defining shared standards at European level;

- Putting in place the necessary technical requirements at national and European level, to address the aspect of legal access to data, that should not be underestimated, by creating a European framework capable of enabling the use of data from commercial sources (e.g. mobile traffic).

In terms of capability, it should be emphasized the need to accompany the above with a strategic investment in the development of new skills starting from basic training, fostering a good understanding by the general public and key stakeholders of the importance of Trusted Smart Statistics for having well-informed citizens and policy makers in a 'datafied' society. This includes raising the awareness among National Statistical Institutes of the importance of their commitment and among decisions makers concerning the use of third-party data for the public interest. Furthermore, the development of European and national research projects with heterogeneous working groups that include thematic, methodological with reference to statistical and technological skills is recommended. The combination of several skills applied to innovative projects will allow the progressive convergence towards the creation of new ”integrated skills” able to fully exploit the potential of new sources.

4

Conclusions

This deliverable has shown that big data or new data sources can be employed to improve the quality of indicator estimation for granular spatial units, and, here, especially for indicators of poverty and well-being. The results can be great examples of the modern production of SDGs indicators that integrates conventional and non-conventional and potentially unstructured data sources. It, thereby, uses data from diverse statistical data processes and applies model- and design-based approaches to inference. However, there are additional steps to be taken in any given application as compared to more traditional methods (the especially tedious tasks of data cleaning and data preparation before being actually able to use the new data only being just one of these).

Further research efforts are needed in order to improve current methods and to facilitate the additional tasks just mentioned and to thereby improve the cost-benefit relation. To mention just some fields: multi-source estimation, meta-analysis, and big data analysis proper. In addition, further research is also needed in other fields that are vital for the measurement of poverty and well-being like regional consumer price indices and their quality. It is important to monitor/estimate indicators in/for the places were people actually live (see scanner data application). Properly taking measurement errors into account in modelling is another broad field that is important across all the topics raised above.

Future research does, in turn, crucially depend on the availability and accessibility of data. For researchers at universities, it is not always easy (if possible at all) to get access to the most promising data sets. Administrative procedures tend to take up large amounts of time. By the same token, costs for certain data sets (e.g. high resolution satellite images) are large, de facto excluding many researchers from bringing their expertise to this area. Easier and less costly data access would therefore be a prerequisite to ensure progress in this fascinating and very promising research area. These points were specifically detailed in Subsection 3.b of this deliverable.

Bibliography

Adelman, I. (1958). A new approach to the construction of index numbers. *The Review of Economics & Statistics 40*, 240–249.

Aitchison, J. (1986). *The statistical analysis of compositional data.* London: Chapman & Hall.

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213. New York, NY: Springer New York.

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package `laeken`. *Journal of Statistical Software 54*(15), 1–25.

Andersson, C., G. Forsman, and J. Wretman (1987a). Estimating the variance of a complex statistic: a Monte Carlo study of some approximate techniques. *Journal of Official Statistics 3*(3), 251–265.

Andersson, C., G. Forsman, and J. Wretman (1987b). On the measurement of errors in the Swedish consumer price index. *Bulletin of the International Statistical Institute 52*, 155–171.

Anselin, L. (1980). Estimation methods for spatial autoregressice structures. *Regional Science Dissertation and Monograph Series, Ithaca: Bornell University.* (8).

Anselin, L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogenety. *Geographical analysis 20*(1), 1–17.

Anselin, L. (1995). Local indicators of spatial assiciantion - lisa. *geographical analysis 27*(2), 93–115.

Anselin, L. (2018). Global spatial autocorrelation (1). Online und `https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#ref-CliffOrd:73`, last visited: 22.09.2020, last updated 04.03.2018.

Anselin, L., I. Syabri, and Y. Kho (2006). Geoda: An introduction to spatial data analysis. *Geographical Analysis 38*(1), 5–22.

Ardilly, P. and F. Guglielmetti (1992). Optimisation de l'échantillon pour le calcul de l'indice de prix à la consommation. *Insee Méthodes 29-30-31*, 71–123.

Arima, S., W. Bell, G. Datta, C. Franco, and B. Liseo (2017). Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society - Series A 180*(4), 1191–1209.

Arima, S., G. Datta, and B. Liseo (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics 42*(2), 518–529.

Articus, C., C. Caratiola, H. Dieckmann, M. Gerhards, R. Münnich, and T. Udelhoven (forthcoming). Measuring well-being at local level using remote sensing and official statistics data. *Rivista di Statistica Ufficiale*.

Balk, B. (1987). Introductory remarks. *Bulletin of the International Statistical Institute 52*, 134–135.

Balk, B. (1989). On calculating the precision of consumer price indices. *Contributed paper for the 47th session of the ISI, Paris*.

Balk, B. (1991). Estimating the precision of a consumer price index; some experiences from the Netherlands. *Contributed paper for the 48th session of the ISI, Cairo*.

Balk, B. and H. Kersten (1986). On the precision of consumer price indices caused by the sampling variability of budget surveys. *Journal of Economic and Social Measurement 14*(1), 19–35.

Banerjee, K. (1956). A note on the optimal allocation of consumption items in the construction of a cost of living index. *Econometrica 24*(3), 294–295.

Banerjee, K. (1959). Precision in the construction of cost of living index numbers. *Sankhyā 21*, 393–400.

Banerjee, K. (1960). Calculation of sampling errors for index numbers. *Sankhyā 22*, 119–130.

Baran, D. and J. O'Donoghue (2002). Price levels in 2000 for London and the regions compared with the national average. *Economic Trends 578*, 28–38.

Baskin, R. (1992). Hierarchical Bayes estimation of variance components for the U.S. consumer price index. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 716–719.

Baskin, R. (1993). Estimation of variance components for the U.S. consumer price index via Gibbs sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 808–813.

Baskin, R. and W. Johnson (1995). Estimation of variance components for the US consumer price index. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 126–131.

Baskin, R. and S. Leaver (1996). Estimating the sampling variance for alternative forms of the U.S. consumer price index. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 192–197.

Bejamin, M., S. Thomas, and T. Suri (2017). There is No Free House: Ethnic Patrinage in a Kenyan Slum. *Working Paper MIT*.

Bell, W. (2012). Notes on a Multivariate Fay-Herriot Model with AR(1) Model Errors.

Berger, Y. G. (1998). Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference 67*(2), 209 – 226.

Bertsimas, D. and M. S. Copenhaver (2018a). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research 270*(3), 931–942.

Bertsimas, D. and M. S. Copenhaver (2018b). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research 270*(3), 931–942.

Biggeri, L. and A. Giommi (1987). On the accuracy and precision of the consumer price indices. methods and applications to evaluate the influence of the sampling of households. *Bulletin of the International Statistical Institute 52*, 137–154.

Biggeri, L. and T. Laureti (2018). Publications, experiments and projects on the computation of spatial price level differences in italy. Technical report, Paper presented at the 3rd Task force meeting the ICP, World Bank held the 27th September 2018, Country case studies: Italy.

BLS (2015). BLS handbook of methods. Technical report, Bureau of Labor Statistics.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*.

Boonstra, H. J. and J. A. van den Brakel (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology 45*(3), 395–425.

Borooah, V., P. McGregor, P. McKee, and G. Mulholland (1996). Cost of living differences between the regions of the United Kingdom. In J. Hills (Ed.), *New Inequalities. The Changing Distribution of Income and Wealth in the United Kingdom*, Chapter 8, pp. 103–112. Cambridge: Cambridge University Press.

Boskin, M., E. Dulberger, R. Gordon, Z. Griliches, and D. Jorgenson (1996). *Toward a more accurate measure of the cost of living.* Advisory Commission to Study the Consumer Price Index.

Bowley, A. (1926). The influence on the precision of index-numbers of correlation between the prices of commodities. *Journal of the Royal Statistical Society 89*(2), 300–319.

Bowley, A. (1928). Notes on index numbers. *The Economic Journal 38*(150), 216–237.

Brown, M., R. de Haas, and V. Sokolov (2018). Regional inflation, banking integration, and dollarization. *Review of Finance 22*, 2073–2108.

Bulman, J., R. Davies, and O. Carrel (2017). Living costs and food survey: Technical report for survey year April 2015 to March 2016. *Newport, Office for National Statistics*.

Bureau of Labor Statistics (2018). *BLS handbook of methods*, Chapter 17. Washington: Bureau of Labor Statistics.

Burgard, J., M. Esteban, D. Morales, and A. Perez (2019). A Fay-Herriot model when auxiliary variables are measured with error. *Test*, 1–30.

Burgard, J., J. Krause, and D. Kreber (2019). Regularized area-level modelling for robust small area estimation in the presence of unknown covariate measurement errors. *Research Papers in Economics 4/19*. Trier University.

Burgard, J. P., M. D. Esteban, D. Morales, and A. PÃ©rez (2020). Small area estimation under a measurement error bivariate Fay-Herriot model. *Statistical Methods & Applications*, 1–30. Online-first version.

Burgard, J. P., J. Krause, and R. MÃ¼nnich (2019). Adjusting selection bias in german health insurance records for regional prevalence estimation. *Population Health Metrics 17*(10), 1–13.

Burgard, J. P. and R. Münnich (2012). Modelling over- and undercounts for design-based Monte Carlo studies in small area estimation: An application to the German register-assisted census. *Computational Statistics & Data Analysis 56*(10), 2856–2863.

Burgess, R., F. Consta, and B. Olken (2012). The Political Economy of Deforestation in the Tropics. *Quarterly Journal of Economics 127*(4), 1707–1754.

Carlin, B. P. and T. A. Louis (2008). *Bayesian methods for data analysis.* CRC Press.

Cavallo, A. and R. Rigobon (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives 30*(2), 151–78.

Chandra, H. and U. Sud (2012). Small area estimation for zero-inflated data. *Communications in Statistics - Simulation and Computation 41*, 632–643.

Chen, M., S. Mao, and Y. Liu (2014). Big data: A survey. *Mobile networks and applications 19*(2), 171–209.

Choudhry, G. and J. Rao (1989). Small area estimation using models that combine time series and cross sectional data. In *Proceedings of Statistics Canada Symposium on Analysis of data in time*, pp. 67–74.

Cliff, A. and J. Ord (1981). *Spatial Processes: Models and Application.* London: Pion.

Daas, P. J., M. J. Puts, B. Buelens, and P. A. van den Hurk (2015). Big data as a source for official statistics. *Journal of Official Statistics 31*(2), 249–262.

Dalén, J. (1995). Quantifying errors in the Swedish consumer price index. *Journal of Official Statistics 11*, 261–276.

Dalén, J. and E. Ohlsson (1995). Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics 13*(3), 347–356.

D'Alò, M., L. di Consiglio, S. Falorsi, and F. Solari (2006). Estimation of variance for consumer price index. In *Proceedings of the Scientific Conference of the Italian Statistical Association.*

Datta, G., P. Lahiri, and T. Maiti (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and inference 102*(1), 83–97.

Davison, A., R. Huser, and E. Thibaud (2013). Geostatistics of dependent and asymptotically independent extremes. *Mathematical Geosciences 45*, 511–529.

De Haan, J., E. Opperdoes, and C. Schut (1999). Item selection in the consumer price index: Cut-off versus probability sampling. *Survey Methodology 25*, 31–42.

Department for Environment, Food and Rural Affairs and Office for National Statistics (2020). Living costs and food survey, 2008-2014. SN: 6385, 6655, 6945, 7272, 7472, 7702 & 7992. [data collection].

Department of Employment (1971). *Proposals for retail prices indices for regions, Cmnd 4749*. London: HM Stationery Office.

Department of the Interior, U.S. Geological Survey (2019). Landsat 8 (L8) Data Users Handbook Version 5.

Destatis (2020a). Mobility indicators based on mobile phone data. `https://www.destatis.de/EN/Service/EXDAT/Datensaetze/mobility-indicators-mobilephone.html` [09.09.2020].

Destatis (2020b). Press release no. 343 of 9 september 2020: Truck toll mileage index, august 2020: +1.2 `https://www.destatis.de/EN/Press/2020/09/PE20_343_421.html` [09.09.2020].

Diewert, W. E. (1995). Axiomatic and economic approaches to elementary price indexes. Technical report, National Bureau of Economic Research.

Dorfman, A., J. Lent, S. Leaver, and E. Wegman (2006). On sample survey designs for consumer price indexes. *Survey Methodology 32*(2), 197.

Duran, H. (2016). Inflation differentials across regions in Turkey. *South East European Journal of Economics and Business 11*, 7–17.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society 22*, 139–153.

Edgeworth, F. (1888). Memorandum by the secretary, Mr F.Y. Edgeworth, on the accuracy of the proposed calculation of index-numbers. *Report of the British Association for the Advancement of Science*, 188–219.

Elteto, O. and P. Koves (1964). On a problem of index number computation relating to international comparison. *Statisztikai Szemle 42*, 507–518.

Èltetö, O. and P. Köves (1964). On an index computation problem in international comparisons. *Statistztikai Szemle 42*, 507–518.

ESA (2020). Copernicus land monitoring service - clc 2012. online: `https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012?tab=metadata`, las checked 10.09.2020).

Esteban, M., M. Lombardía, and E. López-Vizcaíno (in press). Small area estimation of proportions under area-level compositional mixed models. *Test*.

Esteban, M., D. Morales, and A. Pérez (2016). Area-level spatio-temporal small area estimation models. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation*, pp. 205–226. Chichester: John Wiley & Sons, Ltd.

Esteban, M., D. Morales, A. Pérez, and L. Santamaría (2011). Two area-level time models for estimating small area poverty indicators. *Journal of the Indian Society of Agricultural Statistics 66*(1), 75–89.

European Space Agency (2020). User guides: Sentinel-2 msi introduction. online `https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi` last accessed:22.09.2020.

Eurostat and OECD (2012). *Eurostat-OECD Methodological Manual on Purchasing Power Parities*. Luxembourg: Publications Office of the European Union.

Faisal, K., A. Shaker, and S. Habbani (2016). Modelling the Relationship between the Gross Domestic Product and Built-Up Area Using Remote Sensing and GIS Data: A Case Study of Seven Major Cities in Canada. *International Journal of Geo-Information 5*(23).

Fatah, K. and S. Ahmed (2012). Variance estimates for price changes in the consumer price index for Kurdistan region of Iraq (January-December, 2009). *Journal of Basrah Researches (Sciences) 38*(4A), 80–91.

Fava, V. (2007). A precisão dos índices de preços. *EconomiA 8*(1), 39–63.

Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*, 269–277.

Fengki, A., K. Notodiputro, and K. Sadik (2020). Bisakah memperoleh statistik indeks harga konsumen tingkat provinsi di Indonesia dengan ketelitian yang lebih baik? (Can provincial level consumer price indices statistics in Indonesia be obtained with better accuracy?). *Seminar Nasional Official Statistics 1*, 297–306.

Fenwick, D. and J. O'Donoghue (2003). Developing estimates of relative regional consumer price levels. *Economic Trends 599*, 72–83.

Fowler, R. (1973). *Further problems of index number construction*, Volume 5 of *Studies in Official Statistics, Research Series*. London: HM Stationery Office.

Frentzen, K. and R. Günther (2017). Korrektur des Antwortausfalls in der Verdiensterhebung 2015. *Wirtschaft und Statistik* (2), 24–42.

Gajewski, P. (2017). Sources of regional inflation in Poland. *Eastern European Economics 55*, 261–276.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis 1*(3), 515–534.

Gelman, A. and D. B. Rubin (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci. 7*(4), 457–472.

Getis, A. and K. Ord (1992). The analysis of spatial assiciation by use of distance statistics. *Geographical Analysis 24*, 189–206.

Ghosh, M. (2020, August). Small area estimation: its evolution in five decades. *Statistics in Transition New Series 21*(4), 1–22.

Gini, C. (1931). On the circular test of index numbers. *Metron 9*(9), 3–24.

Glaeser, E. (2008). *Cities, agglomeration and spatial equilibrium.* Oxford: Oxford University Press.

Gooding, P. (2016). Consumer price inflation: The 2015 basket of goods and services. Technical report, Office for National Statistics.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Gosh, Tilottama, Sutton, Paul, Elvidge, Christopher, Powell, Rebecca, Baughm, and Kimberly (2010). Shedding Light on the Global Distribution of Economic Activity. *The Open Geograpohy Journal 3*(1).

Harms, A. and S. Spinder (2019). A comprehensive view of machine learning techniques for cpi production. Technical Report Discussion Paper, Statistics Netherlands, The Hague.

Haworth, M. (1996). Re-engineering data production and measuring quality in the UK retail prices index. In *Proceedings: 1996 Annual Research Conference and Technology Interchange*, Washington D.C., pp. 274–302. US Bureau of the Census.

Hayes, P. (2005). Estimating UK regional price indices, 1974–96. *Regional Studies 39*, 333–344.

Heering, S. G., B. T. West, and P. A. Berglund (2017). Applied survey data analysis. *CRC press*.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

ILO, IMF, OECD, UNECE, Eurostat, and The World Bank (2004). *Consumer price index manual.* Geneva: International Labour Office.

Istat (2009). *La misura della povertï¿½ assoluta.* Roma, Italy: Metodi e Norme.stat, Italian national statistical office.

Istat (2010). *La differenza nel livello dei prezzi al consumo tra i capoluoghi delle regioni italiane.* Roma, Italy: Istat, Italian national statistical office.

Jean, N., M. Burke, M. Xie, M. Davis, D. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *SCIENCE 353*(6301), 790–794.

Jevons, W. (1869). The depreciation of gold. *Journal of the Statistical Society of London 32*, 445–449.

Johnson, D. (1975). The accuracy of regression based cost indices. *Journal of the Royal Statistical Society: Series A 138*(3), 411–422.

Kann, K. (2018). Der Einfluss des Mindestlohns auf die Verdienststrukturen. *Wirtschaft und Statistik* (5), 44–56.

Kersten, H. (1985). Nonresponse assessment of a consumer price index. *Journal of Business & Economic Statistics 3*(4), 336–343.

Klick, J. and O. Shoemaker (2019). Measures of variance across CPI populations November 2019. *Proceedings of the Joint Statistical Meetings*, 2242–2253.

Koop, J. (1986). Estimating variance of a consumer price index and some comments on inference. *Journal of Official Statistics* (2), 74–76.

Kosfeld, R., H.-F. Eckey, and M. Schüßler (2009). Ökonometrische Messung regionaler Preisniveaus auf der Basis örtlich beschränkter Erhebungen. Technical Report 33, German Council for Social and Economic Data (RatSWD) Research Notes.

Kott, P. (1983). Estimating the variances of price indexes by half sampling: why it works. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 349–354.

Kott, P. (1984). A superpopulation theory approach to the design of price index estimators with small sampling biases. *Journal of Business & Economic Statistics 2*(1), 83–90.

Kruskal, W. and L. Telser (1960). Food prices and the Bureau of Labor Statistics. *The Journal of Business 33*(3), 258–279.

Laureti, T., C. Ferrante, and B. Dramis (2017). Using scanner and cpi data to estimate italian sub-national ppps. In *Proceeding of 49th Scientific Meeting of the Italian Statistical Society*, pp. 581–588.

Laureti, T. and F. Polidoro (2017). Testing the use of scanner data for computing sub-national purchasing power parities in italy. In *Proceeding of 61st ISI World Statistics Congress, Marrakech.*

Laureti, T. and D. Rao (2018). Measuring spatial price level differences within a country: Current status and future developments. *Estudios de economia aplicada 36(1)*, 119–148.

Leaver, S. (1990). Estimating variances for the US consumer price index for 1978-1986. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 290–295.

Leaver, S. and R. Cage (1997). Estimating the sampling variance for alternative estimators of the US consumer price index. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 740–745.

Leaver, S., J. Johnstone, and K. Archer (1991). Estimating unconditional variances for the US consumer price index for 1978-1986. In *Proceedings of the Survey research methods section, American statistical association*, pp. 614–619.

Leaver, S. and W. Larson (2001). Estimating variances for a scanner-based consumer price index. In *Proceedings of the Government Statistics Section, American Statistical Association.*

Leaver, S. and W. Larson (2002). Assessing the impact of imputation on the sampling variance of the U.S. consumer price index. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2013–2107.

Leaver, S. and W. Larson (2003). Estimating components of variance of price change from a scanner-based sample. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2318–2325.

Leaver, S. and D. Swanson (1992). Estimating variances for the US consumer price index for 1987–1991. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 740–745.

Leaver, S. and R. Valliant (1995). Statistical problems in estimating the US consumer price index. In B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (Eds.), *Business Survey Methods*, pp. 543–566. New York: Wiley.

LeSage, J. and K. R. Pace (2014). What regional scientists nedd to know about spatial econometrics. *The Review of Regional Studies 44*(1), 13–32.

Luna, A., L.-C. Zhang, A. Whitworth, and K. Piller (2015). Small area estimates of the population distribution by ethnic group in england: a proposal using structure preserving estimators. *Statistics in Transition 16*(4), 585–602.

Macarot, P. and F. Statescu (2017). Comparison of ndbi and ndvi as indicators of surface urban heat island effect in landsat 8 imagery: A case study of iasi. *Present Environment and Sustainable Development 11.*

Manski, C. (1993b). Identification of endogenous social effects: The reflection problem. *The review of economic studies 60*(3), 531–542.

Marchetti, S., G. Bertarelli, L. Biggeri, G. Giusti, M. Pratesi, and F. Schirripa-Spagnolo (2019). Small area poverty indicators adjusted using local price indexes. *Italian Conference on Survey Methodology (ITACOSM), 5-7 June, 2019, Florence*.

Marchetti, S., C. Giusti, and M. Pratesi (2016). The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. *AStA Wirtsch Sozialstat Arch 10*, 79–93.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics 31*(2), 263–281.

Marchetti, S. and L. Secondi (2017). Estimates of household consumption expenditure at provincial level in Italy by using small area estimation methods: "real" comparisons using purchasing power parities. *Social Indicators Research 131*, 215–234.

Marhuenda, Y., I. Molina, and D. Morales (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics & Data Analysis 58*, 308–325. The Third Special Issue on Statistical Signal Extraction and Filtering.

Mayhew, M. and G. Clews (2016). Using machine learning techniques to clean web scraped price data via cluster analysis. *Survey Methodology Bulletin 75*, 24–41.

McCarthy, P. (1961). Sampling considerations in the construction of price indexes with particular reference to the united states consumer price index. In G. Stigler (Ed.), *The Price Statistics of the Federal Government*, pp. 197–232. Washington D.C.: National Bureau of Economic Research.

Molina, I. and Y. Marhuenda (2015, jun). sae: An R package for small area estimation. *The R Journal 7*(1), 81–98.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biomtrika 37*(1), 17–23.

Morgenstern, O. (1963). *On the Accuracy of Economic Observations* (2nd ed.). Princeton: Princeton University Press.

Moulton, B. R. (1996). Bias in the consumer price index: what is the evidence? *Journal of Economic Perspectives 10*(4), 159–177.

Mudgett, B. (1951). *Index numbers*. New York: Wiley.

Myklatun, K. (2019). Utilizing machine learning in the consumer price index. Nordic Statistical Meeting, Helsinki, 26-28 August 2019.

Nagayasu, J. (2011). Heterogeneity and convergence of regional inflation (prices). *Journal of Macroeconomics 33*, 711–723.

Nappi-Choulet Pr, I. and T.-P. Maury (2009). A spatiotemporal autoregressive price index for the Paris office property market. *Real Estate Economics 37*(2), 305–340.

Norberg, A. (2004). Comparison of variance estimators for the consumer price index. In *Eighth Meeting of the International Working Group on Price Indices (Ottawa Group), 23–25 August 2004*.

Noč Razinger, M. (2018). Surveys on prices at the statistical office of the republic of slovenia. In B. Lorenc, P. Smith, B. M., G. Haraldsen, D. Nedyalkova, L.-C. Zhang, and T. Zimmermann (Eds.), *The Unit Problem and Other Current Topics in Business Survey Methodology*, Chapter 18, pp. 253–266. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

O'Donoghue, J. (2017). The effect of variance in the weights on the CPI and RPI. *Survey Methodology Bulletin 77*, 1–27.

O'Neill, R., J. Ralph, and P. Smith (2017). *Inflation - History and Measurement*. Basingstoke: Palgrave MacMillan.

ONS (2011). UK relative regional consumer price levels for goods and services for 2010. Technical report, Office for National Statistics, Newport.

ONS (2013). The feasibility of producing regional household final consumption expenditure, UK: 2016. Technical report, Office for National Statistics, Newport.

ONS (2016a). Assessing the impact of methodological improvements on the consumer prices index. Technical report, Office for National Statistics, Newport.

ONS (2016b). The feasibility of producing regional household final consumption expenditure, uk: 2016. Technical report, Office for National Statistics, Newport.

ONS (2018a). Development of regional household expenditure measures. Technical report, Office for National Statistics, Newport.

ONS (2018b). Relative regional consumer price levels of goods and services, UK: 2016. Technical report, Office for National Statistics, Newport.

ONS (2019). *Consumer Prices Index Technical Manual*. Newport: Office for National Statistics.

ONS (2020a). Automated classification of web-scraped clothing data in consumer price statistics. Technical report, Office for National Statistics, Newport.

ONS (2020b). Consumer price inflation item indices and price quotes. data collection.

ONS (2020c). Consumer price inflation, updating weights: 2020. Technical report, Office for National Statistics, Newport.

ONS (2020d). Measures of owner occupiers' housing costs, UK: January to March 2020. Technical report, Office for National Statistics.

Payne, C. (2017). Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices. Technical report, Office for National Statistics, Newport.

Pfefferman, D. (2002). Small area estimation - new developments and directions. *International Statistical Review 70*(1), 55–76.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28*, 40–68.

Pfeffermann, D., B. Terryn, and F. A. Moura (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology 34*(2), 235–249.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Powell, B., G. Nason, D. Elliott, M. Mayhew, J. Davies, and J. Winton (2018). Tracking and modelling prices using web-scraped price microdata: towards automated daily cpi forecasting. *Journal of the Royal Statistical Society, Series A 181*, 737–756.

Prasad, N. and J. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association 85*(409), 163–171.

Purwono, R., M. Yasin, and M. Mubin (2020). Explaining regional inflation programmes in Indonesia: Does inflation rate converge? *Economic Change and Restructuring*, 1–20.

R Core Team (2019a). R: a language and environment for statistical computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2019b). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ralph, J., R. O'Neill, and P. Smith (2020). *The Retail Prices Index: a Short History.* Cham, Switzerland: Palgrave Pivot.

Rao, D. P. and G. Hajargasht (2016). Stochastic approach to computation of purchasing power parities in the international comparison program (icp). *Journal of econometrics 191*(2), 414–425.

Rao, J. and I. Molina (2015a). *Small area estimation* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K. and I. Molina (2015b). *Small Area Estimation.* Hoboken, New Jersey: John Wiley & Sons.

Rao, S. R., B. I. Graubard, C. H. Schmid, S. C. Morton, T. A. Louis, A. M. Zaslavsky, and D. M. Finkelstein (2008). Meta-analysis of survey data: application to health services research. *Health Services and Outcomes Research Methodology 8*(2), 98–114.

Reed, S. and D. Rippy (2012). Consumer price index data quality: how accurate is the US CPI? *Beyond the Numbers 1*, 1–8.

Renwick, T., B. Aten, E. Figueroa, and T. Martin (2014). Supplemental poverty measure: A comparison of geographic adjustments with regional price parities vs. median rents from the american community survey. Technical report, Bureau of Economic Analysis.

Rienzo, C. (2017). Real wages, wage inequality and the regional cost-of-living in the UK. *Empirical Economics 52*, 1309–1335.

Roberts, G. and D. Binder (2009). Analyses based on combining similar information from multiple surveys. In *Survey Research Methods Section of the Joint Statistical Meetings (JSM)*, pp. 2138–2147.

Sammut, F. (2016, 9). *Using generalized linear models to model compositional response data.* Ph. D. thesis, Department of Statistics, University of Warwick.

Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model assisted survey sampling.* Springer Science & Business Media.

Scealy, J. L. and A. H. Welsh (2017). A directional mixed effects model for compositional expenditure data. *Journal of the American Statistical Association 112*(517), 24–36.

Schenker, N. and T. E. Raghunathan (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in medicine 26*(8), 1802–1811.

Shoemaker, O. (2002). Estimation and analysis of variance components for the revised CPI housing sample. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 3208–3212.

Shoemaker, O. (2003). Estimation and comparison of chained CPI-U standard errors with regular CPI-U results (2000-2001). In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 3847–3853.

Shoemaker, O. (2009). The impact of high variances at the lowest aggregate levels on the CPI's all-US-all-items variance. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2835–2843.

Shoemaker, O. and F. Marsh (2011). Revising replicate selection in the CPI variance system. *Proceedings of the Government Statistics Section, Joint Statistical Meetings*, 4068–4079.

Silva-Fernández, L. and L. Carmona (2019). Meta-analysis in the era of big data. *Clinical Rheumatology 38*, 2027–2028.

Skinner, C. (2015). Cross-classified sampling: some estimation theory. *Statistics & Probability Letters 104*, 163–168.

Smith, P. and B. Lorenc (2020, in press). Robust official business statistics methodology during covid-19-related and other economic downturns. *Statistical Journal of the IAOS in press.*

Statistics Bureau of Japan (2020). Regional CPIs for Japan. Technical report, Statistics Bureau of Japan.

Statistisches Bundesamt (2018). Preise Verbraucherpreisindex für Deutschland Qualitätsbericht. Technical report, Statistisches Bundesamt, Wiesbaden.

Statistisches Bundesamt (2020). Regional CPIs for Germany. Technical report, Statistisches Bundesamt, Wiesbaden. search "61111-0011".

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface 14*(127), 20160690.

Suits, D. (1984). Dummy variables: Mechanics v. interpretation. *Review of Economics and Statistics 66*, 177–180.

Szulc, B. (1964a). Index numbers of multilateral regional comparisons. *Przeglad Statysticzny 3*, 239–254.

Szulc, B. (1964b). Indices for multiregional comparisons. *Przeglad statystyczny 3*, 239–254.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tighe, E., D. Livert, M. Barnett, and L. Saxe (2010). Cross-survey analysis to estimate low-incidence religious groups. *Sociological Methods & Research 39*(1), 56–82.

Tillmann, P. (2013). Inflation targeting and regional inflation persistence: Evidence from Korea. *Pacific Economic Review 18*, 147–161.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography 46(supplement).*

Tzavidis., N., L. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society - Series A 181*(4), 927–979.

UK Statistics Authority (2016). Statistics on consumer price inflation including owner occupiers' housing costs - Assessment Report 322. Technical report, UK Statistics Authority, London.

UN (2018). *European Global Navigation Satellite System and Copernicus: Supporting the Sustainable Development Goals - BUILDING BLOCKS TOWARDS THE 2030 AGENDA.*

Valliant, R. (1991). Variance estimation for price indexes from a two-stage sample with rotating panels. *Journal of Business & Economic Statistics 9*(4), 409–422.

Valliant, R. (1992). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics 8*(4), 433–444.

Valliant, R. (1999). Uses of models in the estimation of price indexes: a review. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 94–102.

Valliant, R. and S. Miller (1989). A class of multiplicative estimators of Laspeyres price indexes. *Journal of Business & Economic Statistics 7*(3), 387–394.

van Loon, K. and D. Roels (2018). Integrating big data in the belgian cpi. Technical report, Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland, 7–9 May 2018.

von Hofsten, E. (1959). Price indexes and sampling. *Sankhyā 21*, 401–403.

Weber, A. and G. Beck (2005). Price stability, inflation convergence and diversity in EMU: Does one size fit all? Technical Report CFS Working Paper No. 2005/30, Goethe University, Center for Financial Studies (CFS), Frankfurt a. M.

Weber, W. (1980). A system of variance estimation for the U.S. consumer price index. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 628–633.

Wilkerson, M. (1964). Measurement of sampling error in the consumer price index: First results. *Proceedings of the Business & Economic Statistics Section, ASA*, 220–230.

Wilkerson, M. (1967). Sampling error in the consumer price index. *Journal of the American Statistical Association 62*(319), 899–914.

Williams, E. (2006). The effects of rounding on the consumer price index. *Monthly Labor Review 129*, 80–89.

Wingfield, D., D. Fenwick, and K. Smith (2005). Relative regional consumer price levels in 2004. *Economic Trends 615*, 36–45.

Wolter, K. (2007). *Introduction to variance estimation.* New York: Springer Science & Business Media.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics 24*(1), 53–78.

Wynne, M. and F. Sigalla (1994). The consumer price index. *Federal Reserve Bank of Dallas Economic Review 2*, 1–22.

Xu, T., T. Ma, C. Zhou, and Y. Zhou (2014). Characterizing Spatio-Temporal Dynamics of Urbanization in China Using Time Series of DMSP/OLS Night Light Data. *Remote Sensing 6*, 7708–7731.

Ybarra, L. and S. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*(4), 919–931.

Yesilyurt, F. and J. Elhorst (2014). A regional analysis of inflation dynamics in Turkey. *The Annals of Regional Science 52*, 1–17.

You, Y. and B. Chapman (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology 32*(1), 97.

You, Y. and J. Rao (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics 30*(1), 3–15.

Yue, W., J. Gao, and X. Yang (2014). Estimation of Gross Domestic Product Using Multi-Sensor Remote Sensing Data: A Case Study in Zhejiang Province, East China. *remote sensing 6*, 7260–7275.

Zah, Y., J. Gao, and S. Ni (2003). Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International Journal Remote Sensing*.

Zah, y., J. Gao, and S. Ni (2010). Use of normalized difference built-up index to map urban built-p areas using a semiautomatic segmentation approach. *Remote Sens. Lett. 24*, 583–594.

Zhang, J., P. M. Atkinson, and M. F. Goodchild (2014). *Scale in Spatial Information and Analysis*. Boca Raton: CRC Press.

Zhang, L.-C. (2010). A model-based approach to variance estimation for fixed weights and chained price indices. In M. Carlson, H. Nyquist, and M. Villani (Eds.), *Official statistics: methodology and applications in honour of Daniel Thorburn*, pp. 149–166. Stockholm University and Statistics Sweden.

Zhang, L.-C. (2018). Big data price index. In B. Lorenc, P. Smith, B. M., G. Haraldsen, D. Nedyalkova, L.-C. Zhang, and T. Zimmermann (Eds.), *The Unit Problem and Other Current Topics in Business Survey Methodology*, Chapter 16, pp. 229–236. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological) 67*(2), 301–320.