Use of remote sensing data to improve spatial resolution of well-being and poverty indicators

CBS
Jan van den Brakel and Joep Burger

Trier University
Charlotte Articus, Christopher Caratiola, and Ralf Münnich

University of Southampton
Angela Luna and Nikos Tzavidis

MAKSWELL

Centraal Bureau voor de Statistiek
Trier University
University of Southampton

Related research

- Green (1957) relates physical data extracted from aerial photographs with socioeconomic data of Birmingham, Alabama
- Lo and Farber (1997) assess quality of life by integrating Landsat data wit U.S. Census data
- Ghosh et al. (2013) estimate indicators of human well-being using night-time lights
- Engstrom et al. (2017) link measures of economic well-being with features derived from high resolution satellite imagery in 1291 villages in Sri Lanka

Agricultural/forestry statistics:
Fay and Herriot (1979) or Wagner, Münnich et al. (2017)

Advantages of satellite data

- High spatial resolution
  - Sub 1 meter (Geoeye-1: 0.41m or 0.0147 arc-sec.)
  - to 1km (DMSP/OLS: 1km or 32 arc-sec.)
- Frequent revisits
  - 5 days: Sentinel 2 MSI
  - 16 days: Landsat 8
- Global coverage

However, can we find satellite information that is relatable to social phenomena like poverty?

Advantages of satellite data

- High spatial resolution
  - Sub 1 meter (Geoeye-1: 0.41m or 0.0147 arc-sec.)
  - to 1km (DMSP/OLS: 1km or 32 arc-sec.)
- Frequent revisits
  - 5 days: Sentinel 2 MSI
  - 16 days: Landsat 8
- Global coverage

However, can we find satellite information that is relatable to social phenomena like poverty?

Statistical opportunities

- Downscaling of available statistical data to smaller areas
  - Spatial information
- Combination of different datasets
  - Census areas, municipalities, postcode areas
- Intermediate updating via nowcasting
  - Temporal change
- New remote sensing-based variables:
  - Proximities and infrastructure, vegetation, degree of urbanization
- Remote sensing errors generally unrelated to classical survey errors

View of change: European exurbs, new silk road

Statistical opportunities

- Downscaling of available statistical data to smaller areas
  - Spatial information
- Combination of different datasets
  - Census areas, municipalities, postcode areas
- Intermediate updating via nowcasting
  - Temporal change
- New remote sensing based variables:
  - Proximities and infrastructure, vegetation, degree of urbanization
- Remote sensing errors generally unrelated to classical survey errors

View of change: European exurbs, new silk road

Possible usage of satellite data

- Satellites systematically record measurements of chemical and physical properties
- Typical spectral signatures allow conclusion on properties of the environment
- Proximity and accessibility
- Combination with flight data and other geo-information

Which satellite measures can be related to social variables?

Possible usage of satellite data

- Satellites systematically record measurements of chemical and physical properties
- Typical spectral signatures allow conclusion on properties of the environment
- Proximity and accessibility
- Combination with flight data and other geo-information

Which satellite measures can be related to social variables?

Research questions

- In the course of the MAKSWELL project we investigated:
  1. Model-based downscaling of grid cells from register data:
     Estimation and imputation from greater grid cells to smaller area cells using modelling approaches with optical satellite data
  2. Household-proportional break-down of survey data:
     Estimation of housing quantities using LiDar information to break down census data to city districts
  3. Small area estimation of wealth in Upazilas in Bangladesh using DHS survey data and remote sensing covariates

Study approaches

What approach might be interesting depends on:

- What is the target variable and how is it scaled?
  - Total $\rightarrow$ Redistribution
  - Proportions/fractions $\rightarrow$ Estimation
- What auxiliary information is available?
  Can we build a model?
- Are we interested in grid cells or administrative areas?
  - Do we have a modifiable area unit problem (MAUP)?
- Are we expecting spatial stationarity or interdependence?
  - Need for spatial components

Problem setting

- Aim: Downscaling of median income for Enschede, Den Haag, and Rotterdam
- Target level: 100 m grid cells
- Available data:
  - Median income on level of 500m grid cells, suppressed for areas with less than 30 inhabitants
  - Auxiliary information from administrative sources on level of 500 m grid cells and 100 m grid cells
  - Remote sensing data-based indicators on level of 500 m grid cells and 100 m grid cells
- Use statistical relationship between observed data to predict median income on target level (statistical downscaling)

Auxiliary information reviewed

- Auxiliary information from official statistics
  - Register data with high explanatory power:
    Average value of houses according to tax register, information on social security benefits, ...
  - Information is suppressed in cells with less than 5 observations
    - Fine-resolution target level: large number of NAs
    - No information whether cell is really empty or information is suppressed
- Remote sensing data
  - Open-source data
  - Complete data set with high spatial and temporal resolution

Grid cells

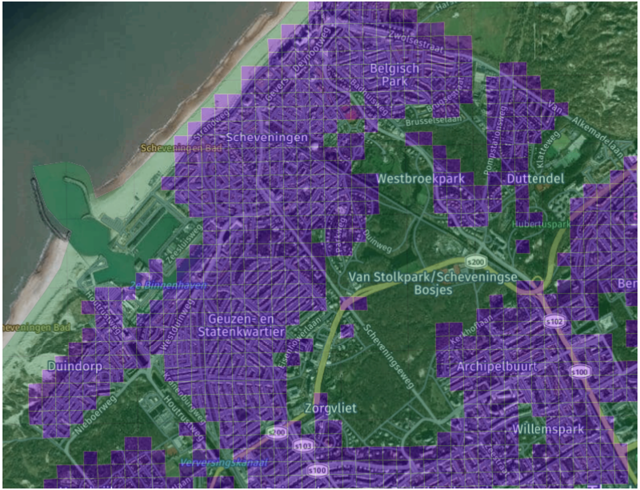


Figure: 500m and 100m grid cells Den Haag

Satellite data

- We calculate 3 indicators from Landsat 8 composites:
- Vegetation density: Normalized Difference Vegetation Index (see Macarot and Statescu 2017)

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

- Ground concealment: Normalized Difference Building Index (see Zah et al. 2013)

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

Satellite data cont.

- Built-up degree (see Faisal et al. 2016)

$$BU = NDBI - NDVI \tag{3}$$

Satellite data used

- Finally, statistics over each 30m pixel within each grid cell are calculated
  - Mean, median, max, min, variance
- We use all images taken by the Landsat 8 satellites in 2017.
- Landsat 8 Collection 1 Tier 1, real-time data raw scenes
- Images are first mosaicked.
- Then, a median filter solves to a composite image for the year.
- We only use images between 1 March 2017 and 31 October 2017.

Spatial structures in cities

- We used the LISA concept by Anselin (1980) to investigate spatial clusters.

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

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

- How similar are the mean differences of a variable X between area i and its neighbours j?
- This allows us to identify clusters of high and low values, based on a permutation test of the null hypothesis of no spatial correlation.

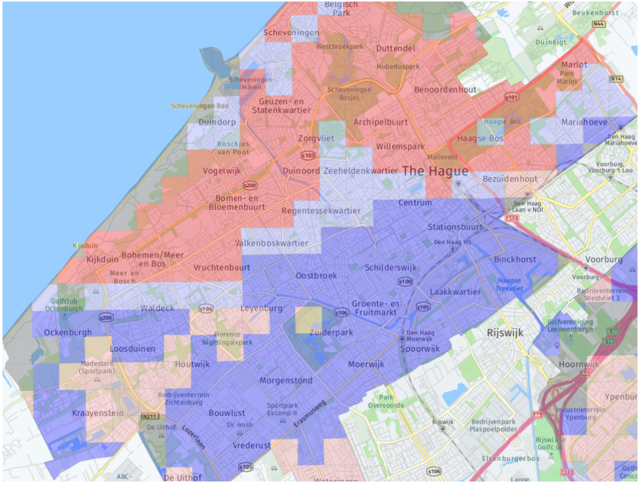


Figure: Local Moran's I of Den Haag for median income

Downscaling

Approach 1: Synthetic approach

Set mean value of smaller cell to value of larger cell it is nested in, i.e.:

$$\overline{Y}_i = Y, \quad i = 1, \ldots, m. \tag{6}$$

(Naive reference approach)

Approach 2: Linear regression model

Fit a suitable linear model on level of larger cells and employ it to predict statistic of interest on target level.

Models I

	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: Model with auxiliary information from administrative sources

Models II

	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: Models using remote sensing data

Results: Den Haag

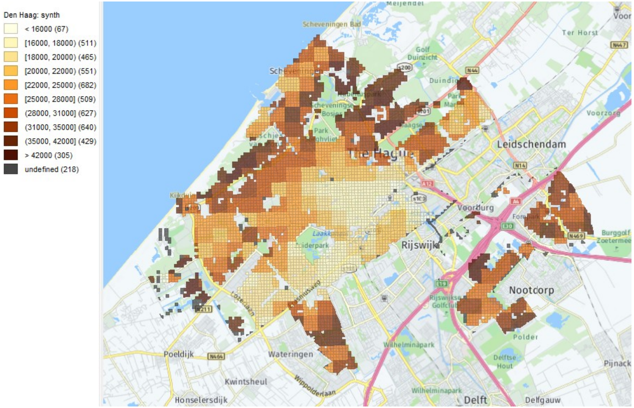


Figure: Downscaling result for Den Haag: Synthetic approach

Results: Den Haag

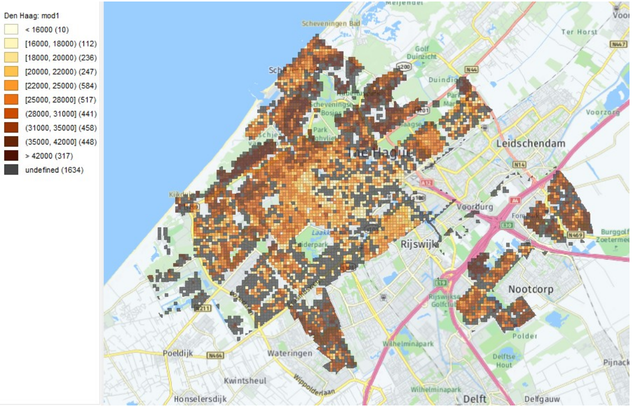


Figure: Downscaling result for Den Haag: Auxiliary information from administrative sources

Results: Den Haag

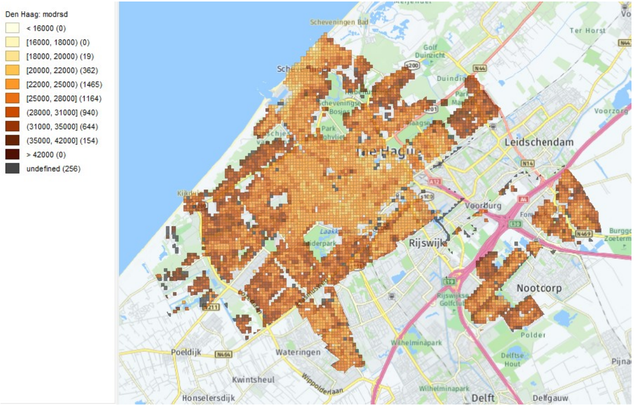


Figure: Downscaling result for Den Haag: Auxiliary information from remote sensing data

SAE with remote sensing data

- Availability, coverage, low cost & high frequency.
- Developing countries. Lack of Census & admin data.
- But, use of remote sensing data e.g. for predicting poverty is complex.
- Model predictive power can be compromised.

SAE with remote sensing data (Cont'd)

- Literature uses mainly supervised learning methods.
- Train machine learning algorithm / statistical model on 'ground truth' (e.g. survey) area data and features extracted from remote sensing data linked to the same geography.
- Use model/algorithm to produce estimates beyond the initial training dataset (downscaling).
- Tempting as one can produce estimates for areas or time periods with scarce, infrequent or completely missing data.

Observations

- The success of using remote sensing data depends on the choice of indicator, the target geography and the availability of ground truth data.
- Model selection and uncertainty estimation are important too.
- However, the choice of prediction method appears to be less important provided it is used appropriately.
- Not always the case. For example, accounting for the sampling design in area-level models in conjunction with the use of automated algorithmic tools (e.g. INLA).
- We illustrate some of these issues in an application.
- Our aim is not to downscale but instead to predict proxy for poverty in formally defined administrative geographies (Upazilas in Bangladesh).

Poverty measurement in Bangladesh

Target: Average WI by Upazila (Level 3).

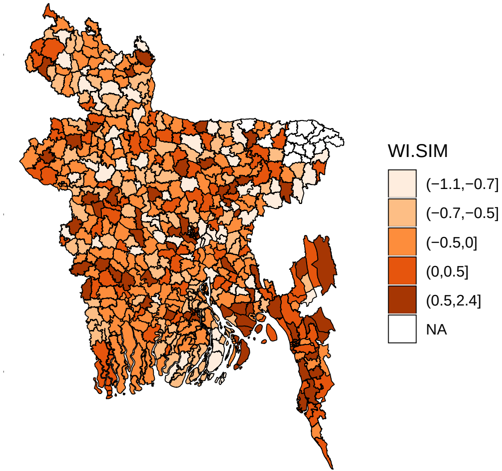
Survey data - DHS 2014

- Stratified 2-stage cluster design. At least one cluster selected in 365/508 (72%) Upazilas
- Response: WI computed via PCA
- n = 17K households. $\bar{n}_i$ = 34 households

RS data:

- 18 variables as starting point in
  - Night time lights
  - Elevation
  - Accessibility to areas with more than 50K people

Direct estimation



Methods

Fay-Herriot model

$$\boxed{\hat{\theta}_i^d = \boldsymbol{x}_i^T \beta + u_i + e_i}$$

$\hat{\theta}_i^d$ a direct estimator of θ_i; $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$.

From a frequentist perspective, σ_i^2 is assumed known.The EBLUP for θ is given by

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

with $\hat{\gamma}_i = \hat{\sigma}_u^2 / (\sigma_i^2 + \hat{\sigma}_u^2)$. Analytic or bootstrap MSE estimation. Available in R packages `emdi`, `sae`.

Methods

Fay-Herriot model

$$\boxed{\hat{\theta}_i^d = \boldsymbol{x}_i^T \beta + u_i + e_i}$$

$\hat{\theta}_i^d$ a direct estimator of θ_i; $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$.

From a Bayesian (HB) perspective:

$$\hat{\theta}_i^d | \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2)$$

$$\theta_i | \beta, \sigma_u^2 \overset{ind}{\sim} N(\boldsymbol{x}_i^T \beta, \sigma_u^2)$$

Information on σ_i^2 can be included via another level or an informative prior.
The posterior mean and variance of θ_i are used for inference.
Available in R packages `BayesSAE`, `hbsae`. BUGS, JAGS, Stan.

Small area models

Point estimates and variances obtained using the sampling design.
Smoothing of variance estimates using GVF.

$$\widehat{WI}_i = \beta_0 + \beta_1 \times ELEV + \beta_2 \times NL + \beta_3 \times ACC + u_i + e_i$$

$u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$. R-INLA latent specification `iid`.

M1 Standard FH model using `sae` with $\sigma_i^2 = \hat{\sigma}_i^2$ treated as known and fixed.

M2 Standard Gaussian model in R-INLA. $\sigma_i^2 = \sigma_e^2$ treated as unknown.

M3 R-INLA with $\sigma_{e_i}^2 = g_i \sigma_e^2$; $g_i = v_i / \bar{v}_i$.

$\tau = 1/\sigma_e^2;\ \pi(\tau) \sim \text{Gamma}\left(\frac{\bar{n}_i - 1}{2} - 1\,,\ \frac{(\bar{n}_i - 1)\bar{v}_i}{2}\right)$.

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

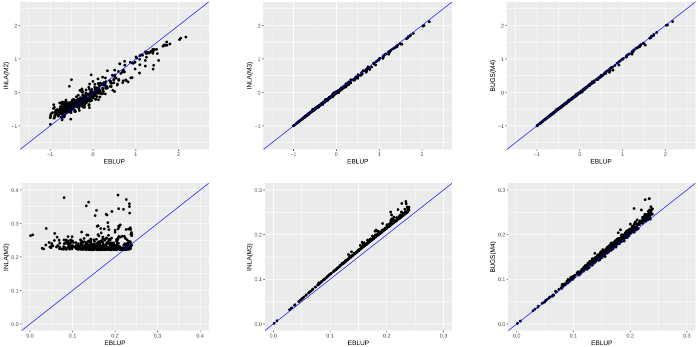
Sensitivity analyses results - Model parameters

- Small differences in the fixed effects
- Large differences in the variance decomposition. Using scaling to allow for heteroscedasticity nearly eliminates all differences

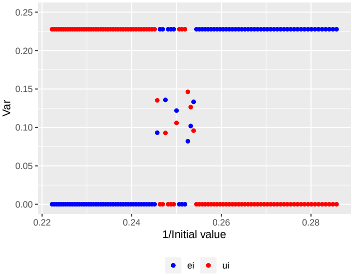
	M1	M2	M3	M4
$\hat{\beta}_0$	0.6662	0.6866	0.6604	0.6469
$\hat{\beta}_{elev}$	-0.0553	-0.0530	-0.0557	-0.0548
$\hat{\beta}_{nl}$	0.3137	0.3112	0.3141	0.3159
$\hat{\beta}_{acc}$	-0.0878	-0.0892	-0.0874	-0.0838
$\hat{\sigma}_e^2$	0.0362	0.1219	0.0423	0.0408
$\hat{\sigma}_u^2$	0.1889	0.1058	0.1800	0.1838

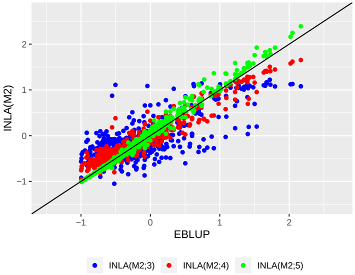
Sensitivity analyses - Point & Uncertainty estimates

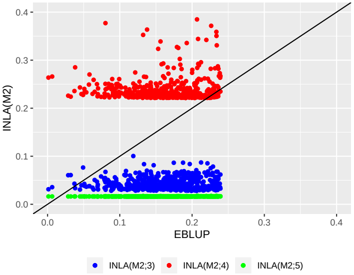
- Some impact on the point estimates
- Large Impact on uncertainty measures



Sensitivity analyses - Initial values







SAE: Discussion and current research

- Ready to use software for approximate Bayesian inference offers interesting possibilities. However, some degree of specialized knowledge and understanding of the model is necessary for its correct use.
- Accounting for the sampling design often overlooked.
- Current work on a similar assessment with spatial models.
- Methods to account for cluster displacement under development.
- Modelling more complex indicators e.g. measuring extreme poverty important but often overlooked. More emphasis on defining appropriate remote sensing covariates is needed in this case.

Downscaling: Discussion and research

- Data from official statistics: Good predictor, but many NAs
- Part of the variability in median income can also be explained by indicators based on remote sensing data.
- Downscaling result: Overall structures are partly reflected, but models with remote sensing data fail to preserve the range of incomes.
- Proximity to parks and density of buildings may determine the quality of the micro-location but on the level of the whole city other characteristics of location are of relevance, too.
- Possible solutions:
  - Composite approaches that use synthetic approach as a baseline and add small-scale heterogeneity based on the model
  - Include additional information about infrastructure and location within the city

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).

Eurostat (2019). Quality of life indicators-measuring quality of life. Retrieved from hiips://ec.europa.eu/eurostat/statistics-explained/index.php?title=Quality_of_life_indicators_-_measuring_quality_of_life&oldid=175684#Framework_for_measuring_quality_of_life.

Gosh, Tilottama, Anderson, Sharolyn J., Elvidge, Christopher D. and Sutton, Paul C. (2013). Using nighttime satellite imagery as a proxy measure of human well-being. Sustainability 5(12), pp. 4988-5019.

Green, Norman E.(1957). Aerial photographic interpretation and the social structure of the city. Photogrammetric Engineering 23, pp. 89-99.

iib-institut.de (2019). Immobilien und Mietpreise in Trier. Retrieved from hiips://www.capital.de/immobilien-kompass/trier .

Lo, C.P. and Faber, Benjamin J. (1997), Integration of Landsat thematic mapper and Census data for quality of life assessment. Remote Sens. environ. 62, pp. 143-157.

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.

OECD (2017). How's Life? 2017: Measuring Well-being, OECD Publishing, Paris.

Stadt Münster (na). Statistik. Zahlen, Daten, Fakten für Münster. Retrieved from hiips://www.stadt-muenster.de/stadtentwicklung/zahlen-daten-fakten.html.

Statistikamt Nord (na). Statistische Daten zu den Stadtteilen Hamburgs. Retrieved from hiips://www.statistik-nord.de/zahlen-fakten/regionalstatistik-datenbanken-und-karten/hamburger-stadtteil-profile-und-interaktive-karten/.

Stiglitz, Joseph, Sen, Armartya and Fitoussi, Jean-Marc (2009). Report by the commission in the measurement of economic performance and social progress. Paris.

Statistisches Bundesamt (2018). Wohnungen je Hektar. Ergebnisse des Zensus am 9. Mai 2011 in Gitterzellen. Retrieved from hiips://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html.

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.

Image Resources

Copernicus (2017). GIO land (GMES/Copernicus initial operations land) High Resolution Layers (HRLs), processed by ESA. published by GIO land team EEA.

Google (2019). Image of the airport Frankfurt-Hahn. Image by DigitalGlobe 2019. Retrieved from Google Maps under hiips://www.google.de/maps/place/Flughafen+Frankfurt-Hahn/@49.9493503,7.2724645,885m/data=!3m1!1e3!4m5!3m4!1s0x47be180426b4755f:0x3db29c6e96461c58!8m2!3d49.9429026!4d7.2629529, last accessed: 31.05.2019.

NASA Worldview (2019). Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast). available from: hiips://worldview.earthdata.nasa.gov . last accessed 31.05.2019. part of the NASA Earth Observing System Data and Information System (EOSDIS).

United States Geological Survey (2019). USGS Landsat 8 Collection 1 Tier 1 TOA Reflectance images available online: hiip://earthexplorer.usgs.gov/ , last accessed: 31.05.2019.

For the presentation of some results we used maps from the following sources:

Information about the ESRI WoldTopo map see hiips://www.arcgis.com/home/item.html?id=30e5fe3149c34df1ba922e6f5bbf808f.

Information about the HERE Hybrid maps see hiips://developer.here.com/documentation/map-tile/dev_guide/topics/example-basemap.html.