Project - Depopulation sensing by integrative knowledge discovery from big data

Report for the UNDP/UNFPA Depopulation Challenge

PopInsight Team: Sanja Brdar, Nastasija Grujić, Nikola Obrenović, Olivera Novović, Predrag Lugonja, Vladan Minić, Željko Bajić, Milica Milovanović, Nevena Rokvić, Vladimir Crnojević

Issued by:	BioSense Institute
Issue date:	22/07/2021
Due date:	31/07/2021

DATA SOURCES **3**
Radio-base station data (Telecom) 4
Activity data (Telecom) 4
Connectivity data (Telecom) 5
Mobility data (Telecom) 5
Land cover/land use data (Copernicus CORINE land cover/land use) 6
Open street maps data 6
Census of population data 6

DATA PROCESSING **7**
Processing activity data 7
Processing connectivity data 7
Processing mobility data 8
Processing land use/land cover data 9
Processing OpenStreetMap data 9

INDICATORS **10**
Extracted indicators 10
Example indicators 13

MACHINE LEARNING **17**
Predictive modeling 17
Indicators importances 18
Model summary explanation 19

RESOURCES **20**
Literature 20

DATA SOURCES

PopInsight project uses several data sources originating from telecom provider, satellite imagery products, open data platforms as well as national statistics data and explores them jointly to detect depopulation trends. Summary of used data sources is provided in Table 1.

Data sources	Size	Spatial resolution	Temporal resolution
Radio Base Stations from Telekom Serbia	3.5 MB - raw data file 5.3 MB - extended with information on municipality and geometry file to assign each antenna to corresponding municipality	antenna	-
Activity data from Telekom Serbia	15 GB	antenna	1 hour
Connectivity data from Telekom Serbia	40 GB	antenna	1 hour
Mobility data from Telekom Serbia	843.9 GB	antenna	No temporal aggregation, records generated in time of SMS, call or internet activity
POI from OpenStreetMaps	214 MB	-	-
Traffic infrastructure from OpenStreetMaps	142 MB	-	-
CENSUS of population for 2011 and estimation for 2019	5 kB	municipality	2011 and 2019
Corine LU/LC	700 MB	100m	2018

Table 1: Data sources used in the project

Radio-base station data (Telecom)

This dataset contains a list of antenna coordinates with their unique identifiers.

- LA - identifier of local area related to base station coverage
- Cell ID - identifier of base station
- Longitude - antenna's longitude
- Latitude - antenna's latitude

By using a combination of Cell ID and LA identifiers, we were able to spatially reference telecom records. The dataset consists of 4519 unique pairs of coordinates and by using it we were able to spatially reference telecom records. This data source is used in the extensive data processing steps performed on activity, connectivity and mobility data.

Activity data (Telecom)

Telecom activity data reflects the amount of telecommunication activity that occurred on a particular antenna. In the dataset different types of activity can be distinguished:

- Call In - activity which is proportional to the amount of received call in a radio base station
- Call Out - activity which is proportional to the amount of generated calls in a radio base station
- SMS In - activity which is proportional to the amount of received SMS in a radio base station
- SMS Out - activity which is proportional to the amount of sent SMS in a radio base station
- Uplink bytes - activity that is proportional to the amount of uploaded bytes in a radio base station using an internet service of telecom provider
- Downlink bytes - activity that is proportional to the amount of downloaded bytes in a radio base station using an internet service of telecom provider
- Number of sessions - activity that is proportional to the number of internet sessions user made in the a radio base station domain using an internet service of telecom provider

Beside mentioned types of activities, the dataset contains following columns:

- LA - identifier of local area related to base station coverage
- Cell ID - identifier of base station
- Date Time - date and hour in which activity is aggregated
- IMSI - the activity is identified by IMSI code, which enables us to distinguish the mentioned type of activities per country code.

Telecom activity was aggregated in a time resolution of one hour. We received the activity dataset for a period of the first 6 months in 2020.

Connectivity data (Telecom)

Connectivity datasets indicate the amount of voice activity exchange between two pairs of radio base stations. It represents a directed graph between the originating and terminating antenna. The weight of the links can be determined by two different indicators:

- Time duration of calls - total duration of exchanged calls between two antennas in predefined time window
- Number of calls - number of exchanged calls between two antennas in predefined time window

In the dataset, links that were initiated or terminated by a user who is using a service from a different telecom operator can be distinguished. The dataset was aggregated in a time resolution of one hour. The dataset contains separate files for each day in the period of the first 6 months in 2020. The dataset schema contains columns:

- Orig LA - identifier of originating local area related to base station coverage
- Orig CellId - identifier of originating base station
- Term LA - identifier of terminating local area related to base station coverage
- Term CellId - identifier of the terminating base station
- Date Time - date and hour in which voice connectivity is aggregated
- Duration Of Calls - aggregated duration of all voice records in time frame in seconds
- Number Of Calls - number of calls that were registered in a time frame

The unique identifier of the antenna is a combination of LA and CellId identifiers due to the specific demands in telecom network infrastructure architecture.

Mobility data (Telecom)

The mobility dataset contains a set of telecommunication records (SMS In/Out, Call In/Out, Internet activity) performed by randomly selected anonymized users for a period of two weeks. Essentially, whenever a user made a telecommunication traffic, the record was created and stored in a telecom database. The dataset contains columns:

- LA - identifier of local area related to base station coverage where user performed an activity
- Cell ID - identifier of base station where user performed an activity
- Time - time when an activity is performed

- Type of activity - type of telecommunication activity, such as SMS In/Out, Call In/Out, and Internet Activity

The dataset was given for the period of 6 months, however, the subset of users was changed every two weeks due to standard privacy protection procedures for such data type. Spatial or temporal aggregation was not applied.

Land cover/land use data (Copernicus CORINE land cover/land use)

Corine LU/LC contains land use data divided in 44 categories (source: https://land.copernicus.eu/eagle/files/eagle-related-projects/pt_clc-conversion-to-fao-lccs3_dec2010). It is provided for 1990, 2000, 2006, 2012, and 2018. As each LU/LC image is a product of different data sources (primarily satellite data), properties of each image are different and can be found on https://land.copernicus.eu/pan-european/corine-land-cover.

Open street maps data

OSM data contains vector-based datasets in which we distinguish:

- Point-based: point of interest, such as restaurants, gas stations, etc.
- Line-based, e.g. road or rivers
- Polygon-based, e.g. buildings, land use

For the purpose of detecting population trends we will utilize points of interest, as well as data about road infrastructure.

Census of population data

Official population data from 2011, as well as estimated population data from 2019 were used. These data are available at the municipalities level. Difference in the population was calculated to serve as a baseline for comparison with indicators derived from all other data sources.

> **Heterogeneous data sources need further processing to extract valuable information.**

DATA PROCESSING

Processing activity data

First, we eliminated data records for which we did not have antenna coordinates. After that spatial and temporal aggregation was performed. We summarized all types of the activities on the municipality level in order to align the resolution of the activity dataset with census data. Temporal aggregation was performed on different levels, such as daily, weekly, and monthly. Attribute IMSI having information on the country codes was kept in aggregated files to enable extraction of the indicators related to the foreign activity.

Processing connectivity data

To explore telecom connectivity data in space and time, we performed additional aggregations over the original dataset. First, we assigned a unique antenna key to each originating and terminating antenna based on a combination of LA and CellId values. Next, we excluded those records from the dataset that have origin or destination outside MTS, because for those records we don't have information about base stations. To obtain georeferenced records we performed joins with geographical data based on unique antenna keys. Geographical data associated with each antenna is latitude and longitude, municipality and centroid point (centroid of spatial polygon of the municipality). Next, we performed additional time and space aggregations. We performed time aggregations on level hour, day period, day and month and spatial aggregations on level base station and municipality.

We obtained eight aggregated data sets that follow the schema:

- connectivity_hour_base_station - aggregated connectivity between base stations on hourly basis
- connectivity_hour_municipality - aggregated connectivity between municipalities on hourly basis
- connectivity_day_period_base_station - aggregated connectivity between base stations on day period level (day period refer to night, morning_and_afternoon and evening depending on hour interval)
- connectivity_day_period_municipality - aggregated connectivity between municipalities on day period level
- connectivity_day_base_station - aggregated connectivity between base station on daily basis
- connectivity_day_municipality - aggregated connectivity between municipalities on daily basis

- connectivity_month_base_station - aggregated connectivity between base stations on monthly basis
- connectivity_month_municipality - aggregated connectivity between municipalities on monthly basis

This type of additional aggregations allow us to explore the data at different “zoom” levels. The most detailed set is of course the aggregation on base station level per hour. After this aggregation we obtained unique weighted links between base stations at each hour, where weights are duration of calls and number of calls. If we want to “zoom out” in space we would go to the municipality level. If we want to “zoom out” in time we would go to day or even month level. Special case is time aggregation on day period level which allows us to explore unique patterns related to traffic oscillation during the day. Each of those data sets are further explored and analysed.

Processing mobility data

The provided data set contains telecom records of individual anonymized users. In the beginning, we clean the data by erasing records without the coordinates. Moreover, users in the data set which had only one record were eliminated. Based on antenna coordinates, we labeled records according to the municipality where they were registered. We reconstructed user’s mobility paths by timely ordering a sequence of locations for each user. Based on users’ movement, a directed graph was calculated for each day at the municipality level.

We further assessed home and work locations for each user in the dataset on the municipality level. As the subset of users changes every 15 days, the algorithm was applied to each data sequence separately. Before anything else, we extracted users’ stay locations by calculating the duration that each user spent in each municipality. If a user was in one municipality at some point in time t1 and observed in another municipality at some point in time t2, we calculated the difference (t2−t1)/2 and assigned the half of time to the first municipality, and another half to the second. We then summed up all time intervals assigned to each municipality. If a user spent at least 10 minutes in some location, we marked it as a stay location. Otherwise, it was marked as a pass-by location and excluded from further analysis. Then the dataset was split into two additional datasets. One contained records corresponding to the working hours (from 7 am to 7 pm), and another one to the night (from 7 pm to 7 am) and weekend records. The first one served us to assess the work location, while the other one to estimate a home location, by extracting the location in which the user spent the most of his time in the time period of 15 days. Estimated home locations for the first 15 days, as well as estimated population trends in 2017 are shown on Figure 1 below.

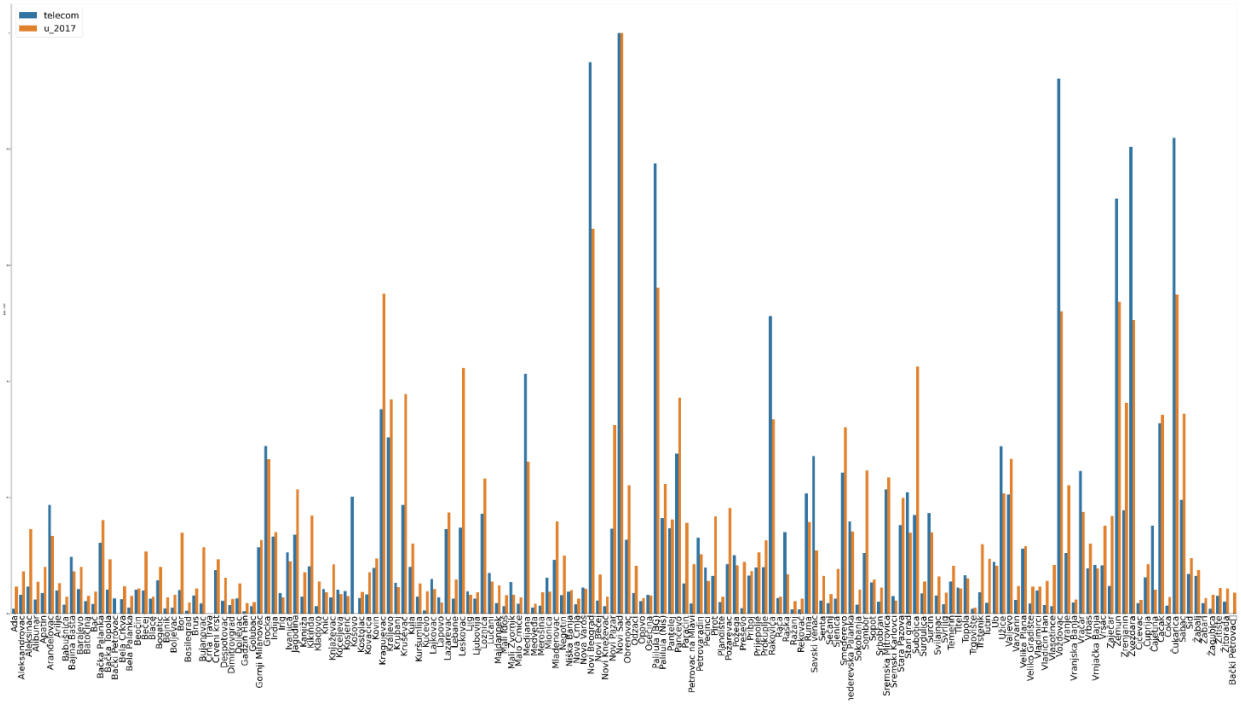

Figure 1: Comparison of extracted home locations with census estimates of the population

Processing land use/land cover data

To calculate the percentage of different land use classes we used the latest Corine satellite-based data product from 2018. Corine dataset was firstly intersected with the shapefile of the municipalities to obtain samples for further calculations. From all available 44 land use categories we used 25 that are present in Serbia.

Processing OpenStreetMap data

Processing of OSM data was applied to each dataset separately, as POI and roads belong to different types of vector features. In order to obtain the reliable distribution of POIs and road traffic, we firstly excluded vector features with invalid geometries. After that, spatially join was performed to enumerate different types of POIs per municipality, as well as to sum up the length of all kinds of roads on municipality level. Some POI categories were excluded where we observed outliers (e.g. POI trees class has an outlier for municipality Užice with more than 10000 annotated trees).

INDICATORS

Activity 3 of the project included extensive extraction of indicators at the level of municipalities. Extracted indicators were further preliminarily analyzed through the scatter diagrams and quantified correlations. Such analysis uses two series of observations, in our case population decline rate and obtained indicator's values across municipalities, and evaluates whether there is a relationship between them. The next subsections report on extracted indicators and provide examples on the explorative part of the analysis. Extracted indicators are inputs into the next project phase - *Data fusion and modelling*.

Extracted indicators

Available data sources (Telecom data, Corine land use/ land cover and OpenStreetMaps) described in the previous report on *Data aggregations, big data processing pipelines* were further processed (summarized, aggregated, statistically analyzed) to extract indicators potentially relevant for explaining and predicting depopulation trends. Overall, more than 150 features were extracted and Table 1 summarizes them.

Data sources	**Indicators**
Activity data from Telekom Serbia	• 5 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during all days in 6 months period, normalized by the latest estimate on the population in municipalities • 5 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during weekdays in 6 months period, normalized by the latest estimate on the population in municipalities • 5 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during weekends in 6 months period, normalized by the latest estimate on the population in municipalities • 5 * 3 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during weekdays in 6 months period during specific period of time (00-07h, 08-15h, 16-24h), normalized by the latest estimate on the population in

	municipalities ● 5 * 3 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during weekends in 6 months period during specific period of time (00-07h, 08-15h, 16-24h), normalized by the latest estimate on the population in municipalities ● 5 indicators extracted from sum of specific type of the activity (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during state of emergency due to COVID-19 pandemic period, normalized by the latest estimate on the population in municipalities ● 5 indicators extracted from sum of the international mobile activities (based on IMSI information) of specific type (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during all days in 6 months period, normalized by the latest estimate on the population in municipalities ● 5 indicators extracted from sum of the international mobile activities (based on IMSI information) of specific type (internet, number of internet sessions, number of calls, duration of calls, number of SMS) during state of emergency due to COVID-19 pandemic period, normalized by the latest estimate on the population in municipalities
Connectivity data from Telekom Serbia	● 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from aggregated connectivity graphs that represent overall voice communication between municipalities during all days in 6 months period ● 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from aggregated connectivity graphs that represent overall voice communication between municipalities during weekdays in 6 months period ● 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from aggregated connectivity graphs that represent overall voice communication between municipalities during weekends in 6 months period ● 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from aggregated connectivity graphs that represent overall voice communication between municipalities during state of emergency due to COVID-19

	pandemic period
Mobility data from Telekom Serbia	• 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from sum of origin-destination flows between municipalities during all days in 6 months period • 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated estimated from sum of origin-destination flows between municipalities during weekdays in 6 months period • 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated estimated from sum of origin-destination flows between municipalities during weekends in 6 months period • 6 local graph properties (degree centrality, pagerank, closeness centrality, eigenvector centrality, betweenness centrality, local clustering coefficient) estimated from sum of origin-destination flows between municipalities during state of emergency due to COVID-19 pandemic period
POI from Traffic infrastructure from OpenStreetMaps	• 5 indicators corresponding to the percentage of overall and specific type of roads (motorway, primary, secondary, tertiary) • 5 indicators related to POIs - their overall number and across general categories (number of commercial sites, office/building space, transport facilities, education)
Corine LU/LC	• 25 indicators corresponding to the percentage of specific land use classes in Corine 2018 maps (Continuous urban fabric, Discontinuous urban fabric, Industrial or commercial units, Road and rail networks and associated land, Port areas, Airports, Mineral extraction site, Dump sites, Construction sites, Green urban areas, Sport and leisure facilities, Non-irrigated arable land, Vineyards, Fruit trees and berry plantations, Complex cultivation patterns, Land principally occupied by agriculture with significant areas of natural vegetation, Broad-leaved forest, Coniferous forest, Mixed forest, Natural grasslands, Transitional woodland-shrub, Inland marshes, Water courses, Water bodies) in municipalities

Table 1: Extracted indicators

We used two measures to estimate the strength of an association between indicators and population decline rate: Pearson's correlation coefficient and Spearman rank correlation.

Pearson's correlation coefficient measures linear association, while Spearman replaces the observations by their ranks and then calculates correlation coefficient that enables capturing nonlinear associations.

Population decline rate is defined as:

$$\frac{population\ estimate\ in\ 2019 - census\ 2011}{population\ estimate\ in\ 2019}$$

and it is multiplied by 100 to express change in percentages. Calculated mean decline rate across municipalities in Serbia is -7.47% and their ranking is available on the link.

The next subsection provides examples of extracted indicators.

Example indicators

Two examples were selected to demonstrate performed analysis on the indicators. Scatter diagrams, with vertical axes representing the indicator while the horizontal corresponding to population decline rate, provide insights into dependence between two variables. Each dot represents one municipality. Furthemore, **the plotly** Python library enabled us to interactively inspect each dot.

The first selected indicator is related to the overall internet during all days in the 6 months period, normalized by the latest estimate on the population in municipalities. This indicator has positive correlation (Pearson 0.305 and Spearman 0.355) with population decline rate (Figure 8) implying moderate dependence between internet consumption and population decline rate. Dot's size correlates with municipality population number.

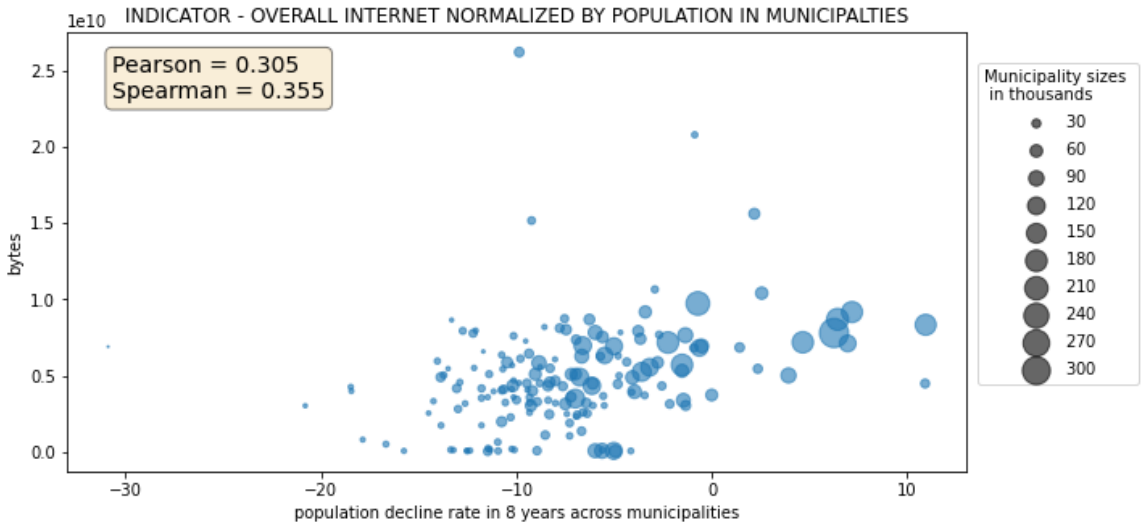

Figure 8: Scatter plot indicator & population decline rate example

Figure 9 is related to the same indicator with an additional color label corresponding to the region of the municipality. Image is available in html format that allows interactive visualization and can be accessed on the link. We can observe on a graph that the municipality Savski venac has the largest internet consumption per estimated number of people. Also municipalities with important tourist locations - Čajetina and Raška stand out in internet usage. Joint modelling of indicators from different sources (mobile phone based and others providing location semantics) enables us to assess their combined effects.

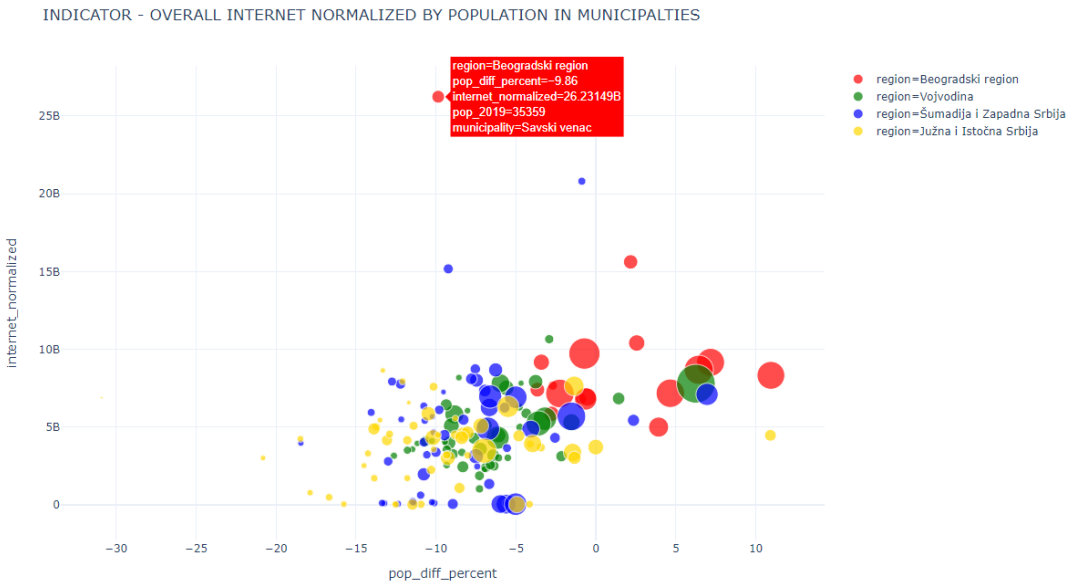

Figure 9: Plotly scatter plot indicator & population decline rate example

Another example is related to the percentage of *Discontinuous urban fabric* indicator derived from Corine land use map from 2018 (Figure 10). *Discontinuous urban fabric* class is assigned to land where features like buildings, roads and artificially surfaced areas occupy significant (range from 30 to 80 % land coverage) surfaces in a discontinuous spatial pattern.

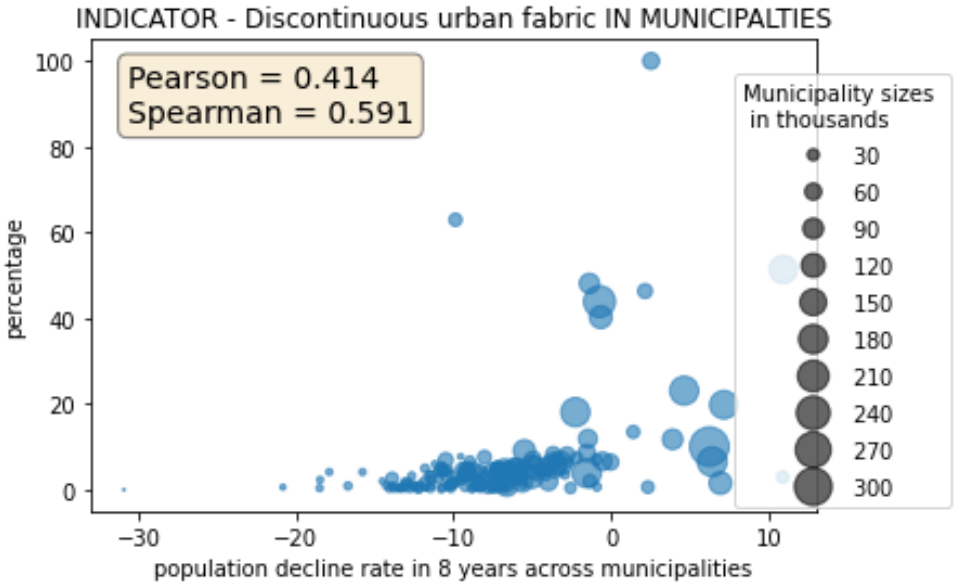

Figure 10: Scatter plot indicator & population decline rate example

Figure 11 is related to the same indicator with an additional color label corresponding to the region of the municipality. Image is available in html format that allows interactive visualization and can be accessed on link. We can observe that municipality Vračar completely belongs to this type of land cover and that the Belgrade region highly overpasses other regions of Serbia in this land use/cover type. Positive correlation implies that municipalities with higher urbanization have smaller population decline rates (in percentages). In-migrations to more urban areas contribute to this trend.

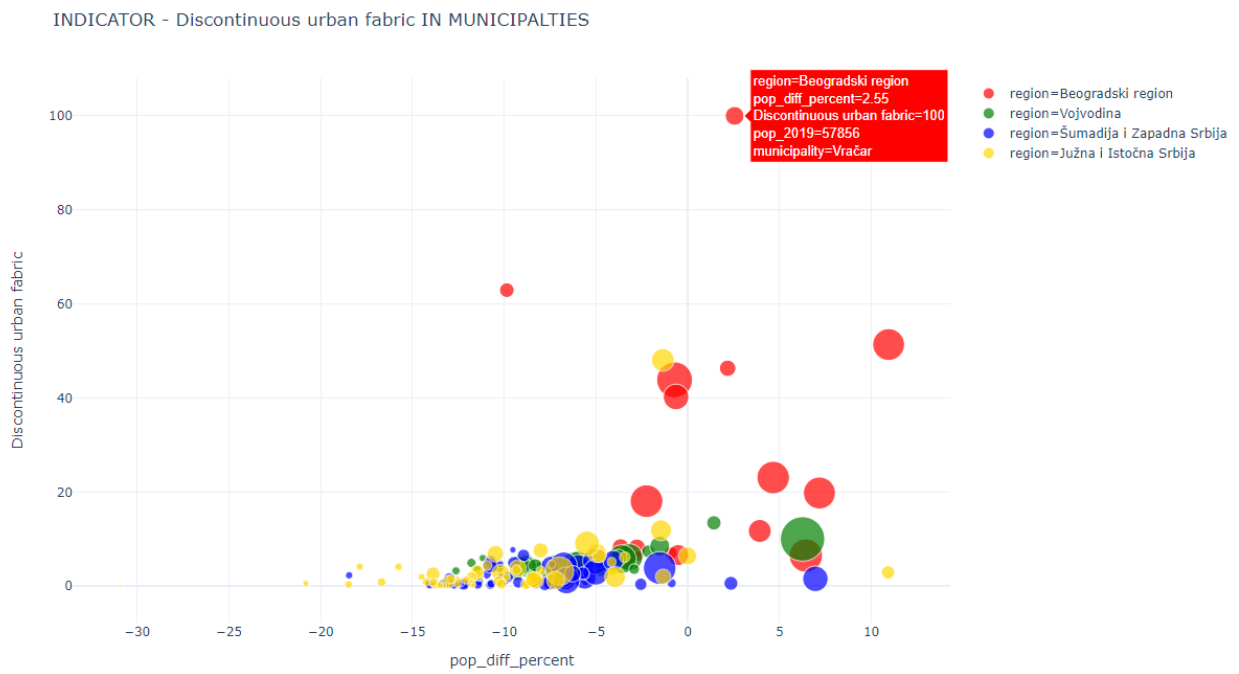

Figure 11: Scatter plot indicator & population decline rate example

All derived indicators are fused in the modeling step, where machine learning based algorithms are utilized to jointly link indicators and population change rates.

MACHINE LEARNING

Predictive modeling

For predictive modeling we utilized a random forest algorithm that is an ensemble method based on decision trees. Decision trees are the smaller models built on subsamples of data/features and the random forest model considers the outputs of the individual trees to make final predictions. Benefit of such an approach is in the reduced risk of overfitting. In regression tasks that we have here in predicting population change, final prediction is the mean prediction of all trees. Figure 12 illustrates the methodology.

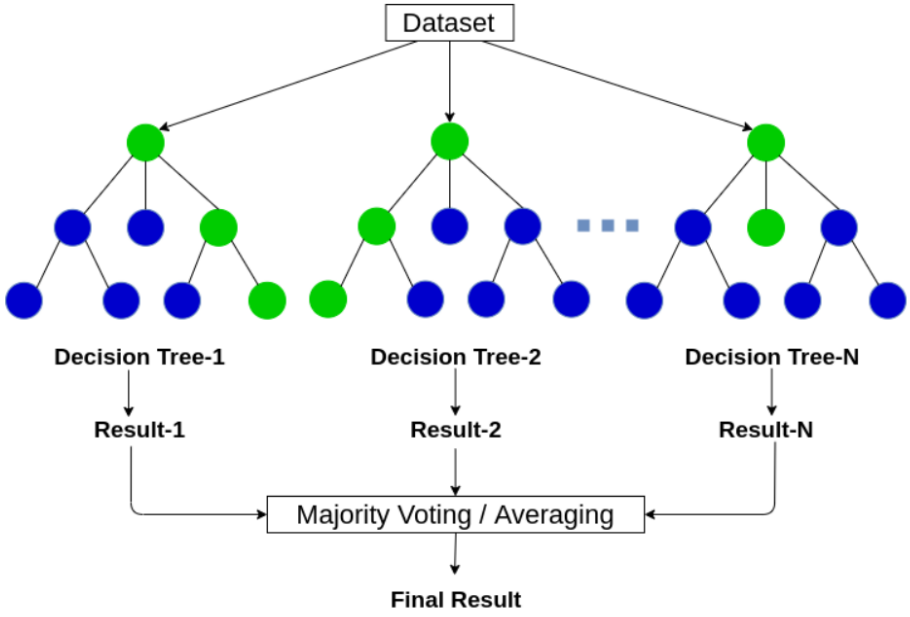

Figure 12: Random forest illustration (source: https://www.analyticsvidhya.com/)

We have utilized machine learning to connect indicators derived from diverse sources of data with population estimates across the municipalities from the Statistical Office of the Republic of Serbia. National population estimates encompass only natural population change determined by births and deaths and official changes of the home addresses. Other changes like emigration abroad or unofficial change of the address remain undetected. Although these estimates have limitations they still capture general trends and can be used as the reference. Through this methodology the main knowledge discovery is in the disagreements between machine learning model and national estimates. This approach unveils to us the municipalities where according to the indicators population estimates are different than expected, i.e. having undetected in-migration or out-migration trends.

Overall 6 random forest models were trained and evaluated: one for each of 5 data sources (activity, connectivity, mobility, corine and OSM) and one for a fused data set. Evaluation was performed in leave-one sample out manner, where sample here represents municipality. The best model is the one learned on all data, and in leave-one-out municipality evaluation procedure achieves correlation with national statistics data of 0.69 (Pearson) and 0.74 (Spearman). When the model is fitted on all data, with parameters tuned through cross validation, agreement between machine learning and national statistics is 0.89 (Pearson) and 0.91 (Spearman).

Indicators importances

Understanding which indicators are the most useful in a model's predictions is crucial and allows us to focus on the most important predictors and lower the complexity of the solution. For a random forest model it makes it easy to evaluate variable contribution to the model and thus measure the variable importance. We used scikit-learn Python implementation of the impurity-based feature importances algorithm, also known as the Gini importance. The higher values denote the more important features. The values are normalized to sum 1, and we provide information in the form of percentages as well.

Indicators' importances results are derived from all 6 models, one for each of 5 data sources (activity, connectivity, mobility, corine and OSM) and one for a fused data set. Before doing importance analysis on fused dataset, a recursive feature elimination algorithm was applied to reduce its size to 30, ensuring that accuracy of the model does not drop and to achieve lower complexity. Recursive feature elimination fits a model and removes the weakest feature and repeats this step until the specified number of features is reached.

Example of feature importance on OSM data is presented in Figure 13. showing ranking of the indicators and that residential road density is the most important indicator from this data source.

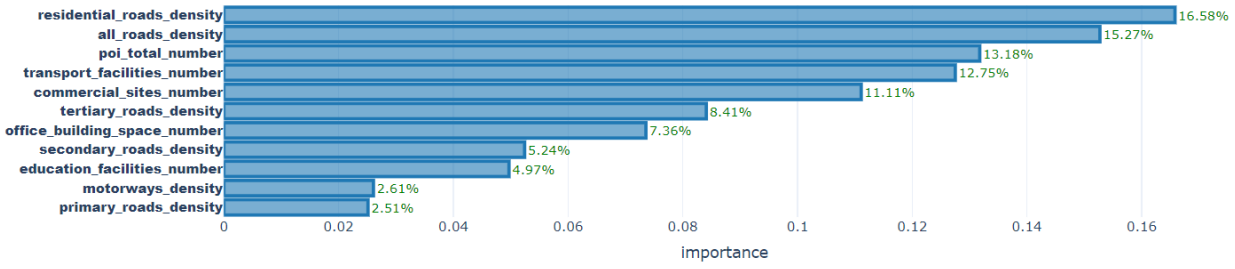

Figure 13: Random forest illustration (source: https://www.analyticsvidhya.com/)

All results on indicators importances are available at PopInsight portal on Results page: https://pop-insight.biosense.rs/results/

Model summary explanation

Further unveiling what model learned can be accomplished with methodology based on SHAP (Shapley Additive exPlanations) that can help in understanding the impact/importance of each feature to model predictions. SHAP is an interpretability method based on Shapley values that are used to calculate how much each individual feature contributes to the model output. SHAP provides a summary view on what the model learned by combining features' importances with features' effects. Methodology can be explained in Figure 14 showing the final result that includes 30 of the most important indicators (features in the machine learning context).

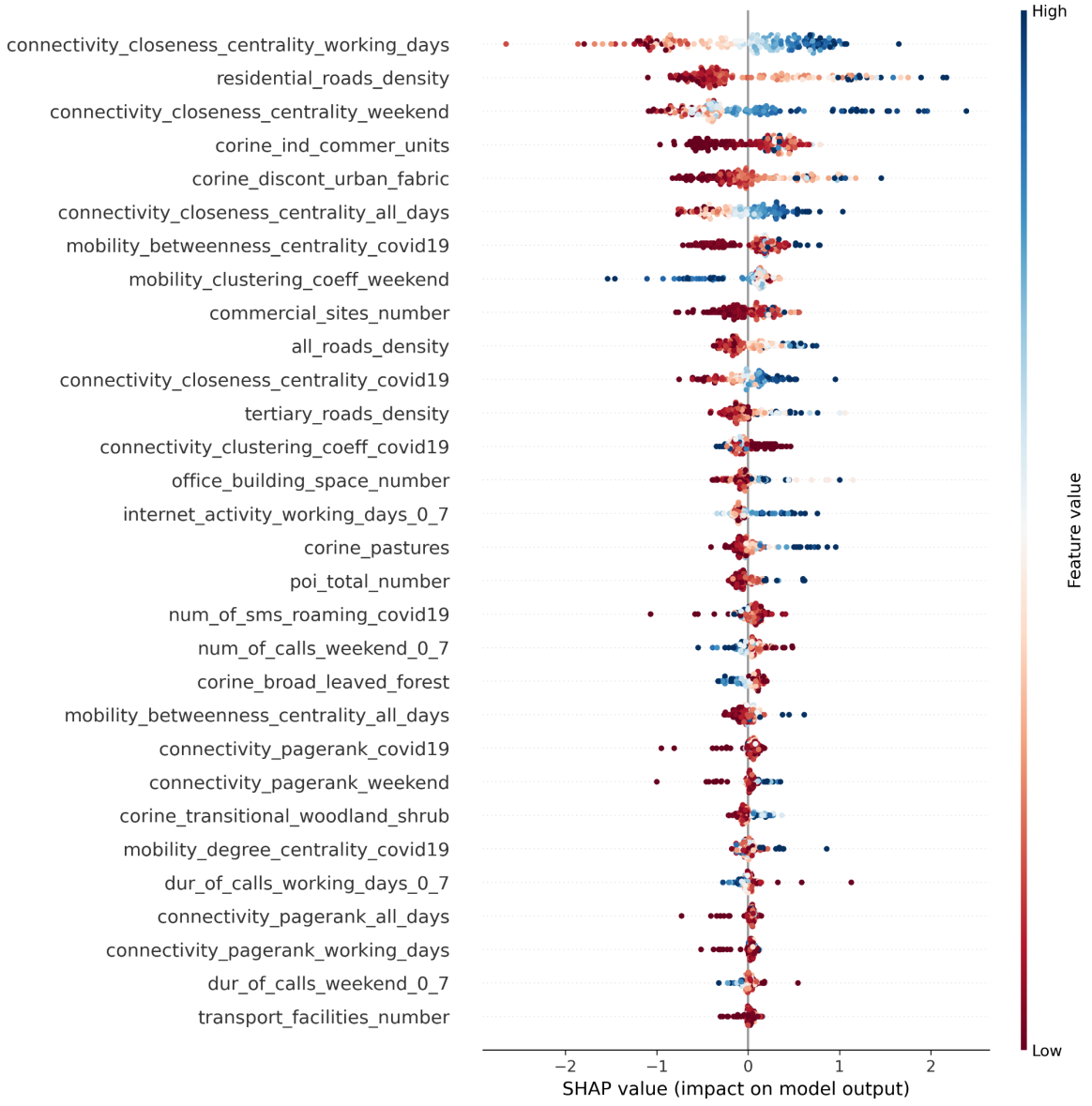

Figure 14: Summary plot of indicators' contribution to the prediction

In our analysis each point on the summary plot is municipality. The position on the y-axis is determined by the indicator and on the x-axis by the SHAP (Shapley value) that indicates its effect. The color represents the value of the indicator (feature) from low to high. Blue dots denote high value of the indicators, while red denote small. SHAP value of zero means that there is no effect of indicator having a value denoted by color to the final prediction. SHAP values >0 increase the predictions, while <0 decrease. In the context of the depopulation we are analyzing here, positive SHAP values indicate better population trends, while negative warns on depopulation. To further illustrate interpretation of the results we can analyze the first ranked indicator - closeness centrality during working days extracted from connectivity data. High closeness centrality itself implies better access to information and more influence on other elements in the network. In our case elements of the network are municipalities connected by telecom traffic. For municipalities with high values of this indicator (blue color) we can expect up to 2% change in population, while municipalities with low values can have 2% loss of the population in the observed period (in our case 8 years). Analysis further continues to the other indicators and as we move to the lower ranked indicators their influence decreases. Through the analysis impacts of indicators sum up to make final predictions. In this way the unveiled model provides us not just a prediction but allows us to have insight into knowledge extracted from the model.

RESOURCES

Literature

[1] Handbook on the use of mobile phone data for official statistics. Report prepared by the Mobile Phone Task Team of the United Nations Global Working Group on Big Data. New York, 2017.

[2] Brdar, S., Novović, O., Grujić, N., González–Vélez, H., Truică, C.O., Benkner, S., Bajrovic, E. and Papadopoulos, A., 2019. Big data processing, analysis and applications in mobile cellular networks. In High-Performance Modelling and Simulation for Big Data Applications (pp. 163-185). Springer, Cham.

[3] Grujić, N., Novović, O., Brdar, S., Crnojević, V. and Govedarica, M., 2019, March. Mobile Phone Data visualization using Python QGIS API. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.

[4] Novović, O., Brdar, S. and Crnojević, V., 2017, April. Evolving connectivity graphs in mobile phone data. In NetMob, The main conference on the scientific analysis of mobile phone datasets (pp. 73-75).

[5] Iqbal, Md Shahadat, Charisma F. Choudhury, Pu Wang, and Marta C. González. "Development of origin–destination matrices using mobile phone call data." Transportation Research Part C: Emerging Technologies 40 (2014): 63-74.

[6] Horanont, Teerayut, Thananut Phiboonbanakit, and Santi Phithakkitnukoon. "Resembling population density distribution with massive mobile phone data." Data Science Journal 17 (2018).

[7] Lai, S., Erbach-Schoenberg, E.z., Pezzulo, C. et al. Exploring the use of mobile phone data for national migration statistics. Palgrave Commun 5, 34 (2019).[1] Brandes, Ulrik. Network analysis: methodological foundations. Vol. 3418. Springer Science & Business Media, 2005.

[8] Networkx documentation, https://networkx.github.io/documentation/stable/

[9] Pappalardo, Luca, Gianni Barlacchi, Filippo Simini, and Roberto Pellungrini. "Scikit-Mobility: An Open-Source Python Library for Human Mobility Analysis and Simulation." NetMob 2019.

[10] Arhipova, Irina, Gundars Berzins, Edgars Brekis, Juris Binde, Martins Opmanis, Aldis Erglis, and Evija Ansonska. "Mobile phone data statistics as a dynamic proxy indicator in assessing regional economic activity and human commuting patterns." Expert Systems 37, no. 5 (2020): e12530

[11] Wang, Zhenzhen, Sylvia Y. He, and Yee Leung. "Applying mobile phone data to travel behaviour research: A literature review." Travel Behaviour and Society 11 (2018): 141-155.

[12] Heiler, Georg, Tobias Reisch, Jan Hurt, Mohammad Forghani, Aida Omani, Allan Hanbury, and Farid Karimipour. "Country-wide mobility changes observed using mobile phone data during COVID-19 pandemic." arXiv preprint arXiv:2008.10064 (2020).

[13] Gibas, Piotr, and Agnieszka Majorek. "Analysis of land-use change between 2012–2018 in Europe in terms of sustainable development." Land 9, no. 2 (2020): 46.

[14] Liu, Xingjian, and Ying Long. "Automated identification and characterization of parcels with OpenStreetMap and points of interest." Environment and Planning B: Planning and Design 43, no. 2 (2016): 341-360.

[15] Breiman Leo, Random Forests, Machine Learning vol.45, pp. 5–32, 2001

[16] https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

[17] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), pp.56-67.