Disaggregation Strategies for Demographic Data in Kazakhstan (CA_POP)

Rakhymbay, Z.[1,2] and Mittlböck, M.[2,3]
[1]Zabira Rakhymbay, Al-Farabi KazNU, Almaty, Kazakhstan, E-mail: zabira.rakhymbay@gmail.com
[2]Paris-Lodron University of Salzburg, Salzburg, Austria
[3]Studio iSPACE, Research Studios Austria, Salzburg, Austria, E-mail: manfred.mittlboeck@researchtudio.at

Abstract
The demographic dynamics are always important for the political, environmental and social-economic forecasts. Because high-resolution of population data is an universal key for solving many challenges of research and analysis. Proper assessment of the demographic situation and its accurate scientific predictions make possible to correct determine and plan the near and distant future changes at the state and local levels. Typically, demographic data collected at a spatially aggregated level. The consequence of this can be misleading results in which the aggregated data refers to a larger spatial unit. The solution to this challenging task can provide spatial disaggregation of population data. The detailed geo-referenced data of population distribution independent from the administrative boundaries is the main advantage of spatial disaggregation. The paper is focusing on the creating spatial disaggregation approach for the different demographic data of Kazakhstan.

1. Introduction

Currently there are more disaggregation approaches as freely available harmonized population grid datasets at the local and global levels. There are the Gridded Population of the World (GPW), the Global Rural Urban Mapping Project (GRUMP), and LandScan Global Population database, EUROSTAT, WorldPOP. It is also necessary to list the World Population Estimated (WPE) dataset from Environmental Systems Research Institute (ESRI). Nevertheless, why is there a need to do additional investigation if there are alternative data as population grids? Essentially, we argue that it is important to have the above-mentioned population grids. However, developing countries, like Kazakhstan, do not have yet their own local population grid data. In addition, the resolution of global grid data is insufficient for activities that are more detailed. Because the resolutions of listed databases are 1 km, only ESRI's WPE data has 230 m at the equator. Secondly, it is necessary to take into account completeness of data for considering study area. Because, population data for Kazakhstan are still poor. In the third, investigation in this research does not consider any data. It goes with data that is sustainable maintained. The chronological scope of this study covers since the period of Kazakhstan's independence in order to have a complete "spatial" picture of the demographic development. Therefore, main and ancillary datasets have chosen with respect to the census years (1989, 1999 and 2009). That means, despite the free availability of data, this study allows for data by time, also by quality. Thus, we brought the idea onto the science community to do spatial disaggregation map for Kazakhstan as additional way of doing population density map.

The main aim of this research is creating more fine-grained population map by using spatial disaggregation strategy for Kazakhstan. The outcomes will be comparable spatially disaggregated population density map for Kazakhstan and spatial temporal story telling application depicting the (long-term) evolution of spatial distribution of population in study area for the years 1989, 1999, 2009. In addition to the expected results in the study will be created geoprocessing model for spatial disaggregation that can be applied at district level or even for another country. As an acronym for spatial disaggregation strategy of creating population density map is chosen "CA_POP" (Central Asian Population). Because after Kazakhstan, this strategy will be used for another Central Asian countries in the future.

2. Disaggregating Approaches

When considering the spatial disaggregation in the literature at the same times there are words such as downscaling, dasymetric mapping, and top-down approach. If consider the definition of words with different scientific papers, then:

«Spatial data disaggregation is an important area of research in GIS and spatial analysis. It involves transferring and decomposing spatial data from one set of a larger spatial unit to a set of smaller spatial units within the same study area» (Flowerdew and Green, 1989).

«Downscaling is a well-known term in environmental studies, which describes the process of generating fine granular data from coarse base data» (Scholz et al., 2013).

«Dasymetric maps are geographical representations based on spatial units in which the target theme (often population density) is distributed as homogeneously as possible. When the spatial units of a dasymetric map are subsets of administrative units with known total population, we can speak of downscaling or disaggregation» (Shu and Lam, 2011).

From these definitions, it follows that the main purpose of these terms eventually get more detailed information on more smaller geographical units. The similarity of the terms it was celebrated on different scientific articles. For example, Scholz et al., (2013) pointed out that term downscaling can be used with the term disaggregation. Bierkens et al., (2000) indicate that terms "upscaling" and "downscaling" can be considered as "aggregation" and "disaggregation" accordingly. In this research, further the terms "spatial disaggregation" or "disaggregation" to be used for the process of creating finer-granularity map.

There are several common methods of modeling population distribution. Many authors have different main categories for grouping disaggregation approach. For example, Deichmann (1996) considers kinds of disaggregation such as an areal interpolation and surface modeling. Wu et al., (2005) grouped general approaches such as an areal interpolation and statistical modeling. Mennis (2009) considers kinds of dasymetric approach such as a traditional cartographic techniques and statistical techniques. Gallego et al., (2011) divided approaches in the following way: estimation method, limit-based methods, geostatistical methods, density maps with a quasi-continuous variation. Referring to the above-mentioned papers, in this study, spatial disaggregation methods have been grouped into two categories depending on whether ancillary information is used. In this research, great attention is paid to the methods of disaggregation approach with ancillary data. Three types of the limit based methods of disaggregation mentioned by Eicher and Brewer (2001) in order to produce population density maps. According to this paper among them the limiting variable method is more accurate than others. In the paper by Li et al., (2007) the three-class method gives less errors than binary method. As mentioned in the paper it relates with an overestimation of residential areas. Because in binary method all population is attributed to one class of land cover and it is urban areas in main cases.

Street weighted method is highlighted in the papers by Xie (1995), Reibel and Bufalino (2005), Mrozinski and Cromley (1999). Generally, this method is used for producing more precise spatially disaggregated population density maps. This method considers that settlements are concentrated in a buffer along a road network. Reibel and Bufalino (2005) notice that street weighted method allows to decrease the errors comparing with areal weighting methods.

Mennis and Hultgren (2006) pointed out another disaggregating method, which is intelligent dasymetric mapping (IDM). According to this paper, IDM specified the relationship of the ancillary classes with the statistical surface. The IDM method implemented as a geographic information system (GIS) extension. For instance, one of them is available for downloading by US Environmental Protection Agency. Another type of statistical techniques is Expectation-Maximization. This method based on areal weighting algorithm. It calculates population density map using maximum likelihood. Gallego and Peedell (2001) developed alternative for this method, it is called as CLC-iterative.

As Li et al., (2007) pointed out that each methods of disaggregation have own strengths and limitations for a desired purpose. Gallego (2010) argues that spatial disaggregation is far from perfect. It have to take into account not only type of disaggregation algorithm but is also the data selecting part. Especially, the quality of the ancillary data is very important, because the end result will depend on directly from it (Steinnocher et al., 2011).

3. Gathering and Organizing Required Data

Miller and Han (2009) pointed out that nowadays is an era of "rich" data and it is important not to be lost among them. Because selecting and collecting data is significant step in any analysis. In addition, unreliable, erroneous, incomplete information can lead to the wrong conclusions and decisions. Therefore, first was done the overview for available datasets which are possible to use in the frame of this research. Despite accuracy of main data, ancillary data sources are also important. Since

census data is distributed along ancillary data and in this case their resolution affect to the final result.

Administrative data for the study area was taken from the Database of Global Administrative Areas (GADM) website (California, 2016), and also was requested from Al-Farabi Kazakh National University (KazNU). The main demographic data was downloaded from official website of Committee on Statistics of Ministry of National Economy of the Republic of Kazakhstan. Committee on Statistics provides the statistics data in all fields. The census data for 2009 covers information since 1959 at the national level. The information for oblast level are given for the last two census: 1999 and 2009 years.

One of the main ancillary data in spatial disaggregation approach is land cover datasets. For this study have been chosen four land cover datasets, such as UMD1993 - University of Maryland (Department of Geography, 1998), GLC2000 (Global Land Cover 2000), GlobCover2009 (UCL, 2010) and GlobeLand30-2010 (NGCC, 2010). The choice of these data was not accidental, because many factors are taken into account. There are:

- Land cover data has to cover the entire study area, i.e. Kazakhstan;
- The data resolution has to be considered. As far as it is high, so accurate and high quality results will be at the end of the analysis;
- Another important factor is accuracy of land cover. Since it may also help to improve the quality of the final results (Mayaux et al., 2006);
- Land cover datasets have been specially selected by the year of census. If consider that the population census in Kazakhstan were held 1989, 1999, 2009, respectively the choice of land cover datasets will be the same or close to these years.

As DEM (Digital Elevation Model) ancillary data was chosen Shuttle Radar Topographic Mission (SRTM) with the resolution 30 m. The data downloaded from Earth Explorer web page (USGS) as 1040 separate tiles in TIFF-format. The Gridded Population of the World (GPW) was downloaded from SEDAC web page (Center for International Earth Science Information Network, 2016). The advantages of GPW are that it uses top-down approach with simple area weighting method for reaggregating data on the administrative boundaries. Additionally it covers data for each 5 years that gives possibility to do time-enabled analysis. Another ancillary data is road networks. It is difficult to find publicly available accurate road networks for Kazakhstan, even for Central Asian territory. Therefore, there were considered VMap0 (Government of Canada, 2000) and OpenStreetMap (OSM) data as alternative sources.

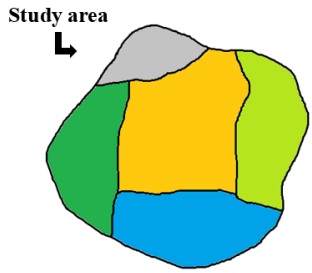
4. "CA_POP" Disaggregation Geoprocessing Model Approach

Core concept of creating "CA_POP" model is based on the "CLC-iterative" method by Gallego and Peedell (2001). The CLC-iterative is explained as regression models which are used to find the relationship between land cover classes and population density. In the paper by Gallego and Peedell (2001) first was used weighting coefficients which provided by the European Environment Agency (EEA) for an aggregated CLC nomenclature. However, it is for all NUTS2 regions. As authors noticed, it is not supply population density for land cover class at the commune level, but it defines median density for each land cover class in each stratum. Thus, the disaggregation method that was used percentage of population in each CLC class was offered. It allows taking into account additional information on the population census data in each class. Therefore, core concept of the method is to assign certain amount of population percentage to each class of land cover datasets (Figure 1).

In this research, initial weighting coefficients do not provide by land cover sources. It is a reason of determining impossibility of population percentage. Therefore, it was applied % of population which are contributed by paper of Gallego and Peedell (2001). As it already mentioned Kazakhstan considered at national and oblast levels. But oblasts differentiate among each other by regional features. Since it was concluded to consider them in two categories (more city-oriented oblasts and more rural based oblasts). Population percentage of EU had been chosen for national level, and % of population of Italy was used for more city-oriented oblasts. Greece's population percentage was assigned to more rural based oblasts.

Since there were used CORINE land cover, it was necessary to do harmonization of classes to GlobeLand30-2010 classification. Because GlobeLand30-2010 was applied as unified classification in this study. Harmonization process had been done according paper by Pérez-Hoyos et al., (2012). They showed matching legends among CORINE and GlobCover2009, GLC2000.

Consequently, there were determined corresponding classes between CORINE and GlobeLand30-2010.



	Classname	Weight for population (W)
	Forest	10%
	Cultivated area	35%
	Artificial surface	50%
	Bareland	5%
	Water bodies	-
		100%

Figure 1: The concept of "CA_POP" disaggregation geoprocessing model

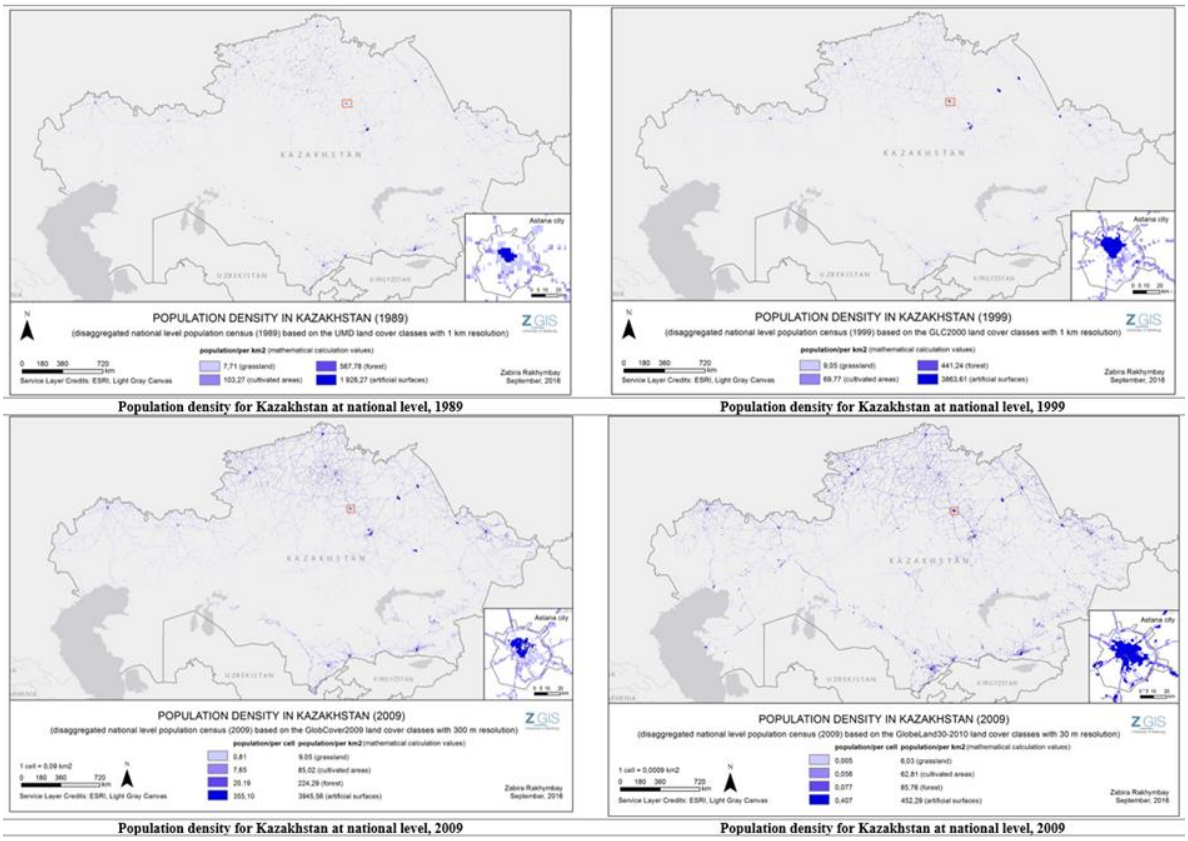


Figure 2: Population density maps for Kazakhstan as the result of CA_POP model

It is important to note that for several classes have been given value 0%. Because they are non-populated areas. Gallego and Peedell (2001) define the artificial surfaces, cultivated areas, grassland and forest as populated areas. Respectively in the remained classes percentage of population are equal to 0. According to land cover classification, which is used in this research, shrubland, bareland, water, wetlands, snow and ice are non-populated classes. The sum of the population percentage should be 100%. After population density is calculated by using population percentage with Equation 1.

$$C_P = \frac{P * W}{100 * S}$$

Equation 1

P – Total population of study area
S – SUMM of cells of study area included in a particular class of land cover
W (%) – % of population attributed to each class of land cover
C_P – number of people in one cell.

5. Results and Discussion

CA_POP disaggregation model was launched for Kazakhstan on two levels for three census years as planned before. From the running workflow was taken population density maps (Figure 2).

There are:

- National level:
  - for the census year 1989 with UMD9293 land cover data;
  - for the census year 1999 with GLC 2000 land cover data;
  - for the census year 2009 with GlobCover2009 land cover data;
  - for the census year 2009 with more finer resolution land cover, which is GlobeLand30-2010.

- Oblast level. In this level, previous order of steps has been saved and model ran for 16 oblasts separately.

Using obtained results of national level it had been calculated attributed population for the following years 1989, 1999, 2009 and compared with real population data from census (Table 1). Table shows population density values for each land cover class and sum of cells of corresponding class. As mentioned previously these classes are defined as populated areas in the paper by Gallego and Peedell (2001). Using values of population density and sum of cells was calculated population count for each class. Then sum of these counts gave attributed total population at national level. Further that the acquired results are almost accurate.

Since main result values as total population give positive result it is possible to go deeper to compare particular region, for instance, Astana (Figure 3). Astana is young capital of Kazakhstan that became the capital since 1998. In that period the city has 319 thousand people and the territory was 258 sq. km. Nowadays Astana occupies 710.2 sq.km and the population increased more than twice (2009). The main reason for the increasing of the population density in a given area is also related to socio-economic factors. It means that people migrate with high intensity in search of good job and high quality of life with good social conditions.

Figure 3 shows the results of CA_POP model for Astana city for last three census years. Also there are illustrated results of GPW for corresponding three years and WPE for 2015. As it can be seen from the figure CA_POP results for 1999 has same resolution with GPW for 2000. But as disaggregate model CA_POP has better results than GPW. In the case of WPE things are differently. The WPE data has better resolution than results than results of CA_POP for 2009. Nevertheless, it should be noted that WPE is available only for one year. Since for monitoring and observation of the demographic dynamics is required data with different time intervals. Accordingly, the choice of this GRID would be erroneous. Also should be noted that there are not so much information about the WPE data.

Table 1: Comparison of attributed and real population

LC class	**Census year: 1989; LC: UMD (1 km)**			**Census year: 1999; LC: GLC2000 (1 km)**		
	SUMM of cells per class	PopDens per class	Population per class	SUMM of cells per class	PopDens per class	Population per class
Cultivated areas	25604	103,27	2644125	34999	69,77	2441880
Forest	600	567,78	340668	713	441,24	314604,1
Grassland	50440	7,71	388892,4	39701	9,05	359294,1
Artificial surfaces	6663	1928,27	12848063	3071	3863,61	11865146
Attributed population			**16221748,49**	**Attributed population**		**14980925**
Real population			**16222324**	**Real population**		**14981281**
LC class	**Census year: 2009; LC: GlobCover2009 (300 m)**			**Census year: 2009; LC: GlobeLand30-2010 (30 m)**		
	SUMM of cells per class	PopDens per class	Population per class	SUMM of cells per class	PopDens per class	Population per class
Cultivated areas	341015	7,65	2608765	46162034	0,056531	2609586
Forest	16655	20,18	336097,9	4355773	0,077185	336200,3
Grassland	472751	0,81	382928,3	70717340	0,005433	384207,3
Artificial surfaces	35707	355,101	12679591	31149029	0,407062	12679586
Attributed population			**16007382**	**Attributed population**		**16009580**
Real population			**16009597**	**Real population**		**16009597**

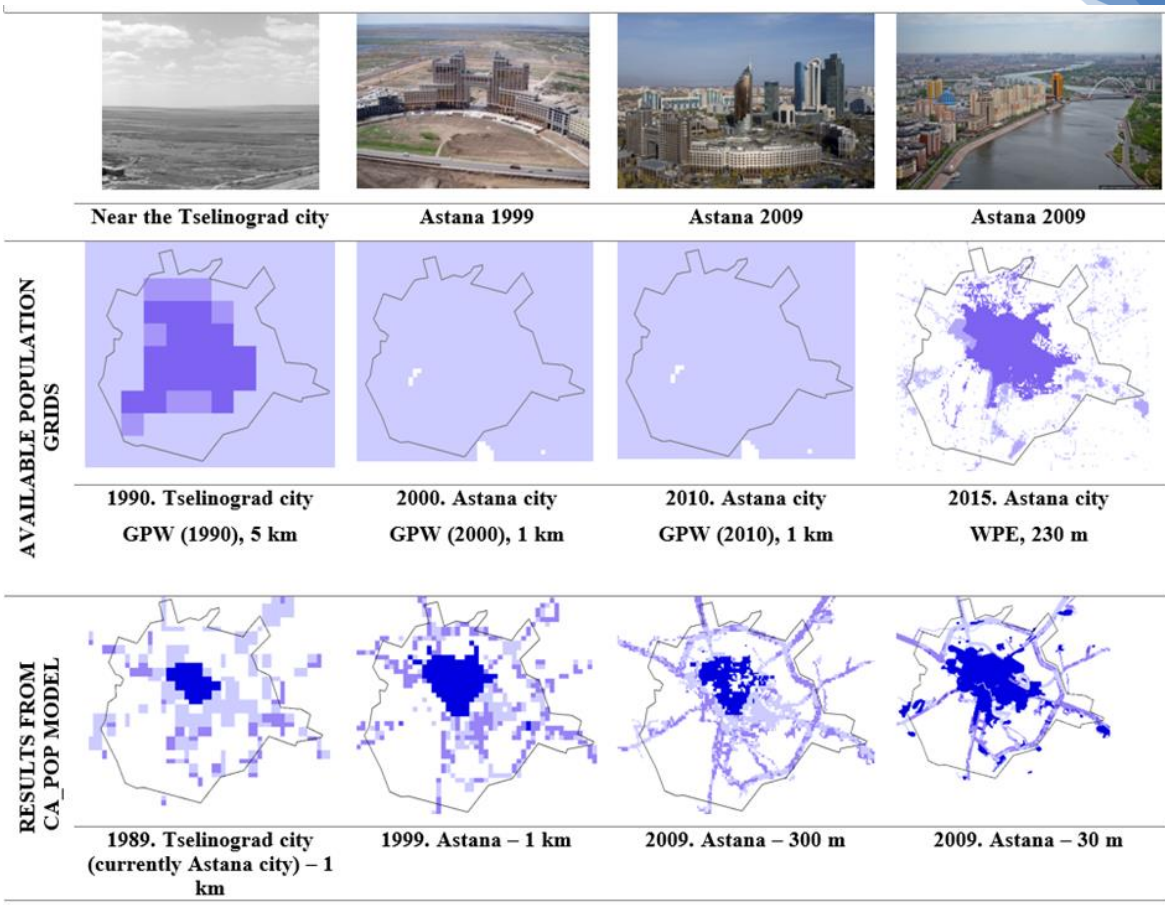


Figure 3: Comparing results of CA_POP model for Astana city with GPW and WPE

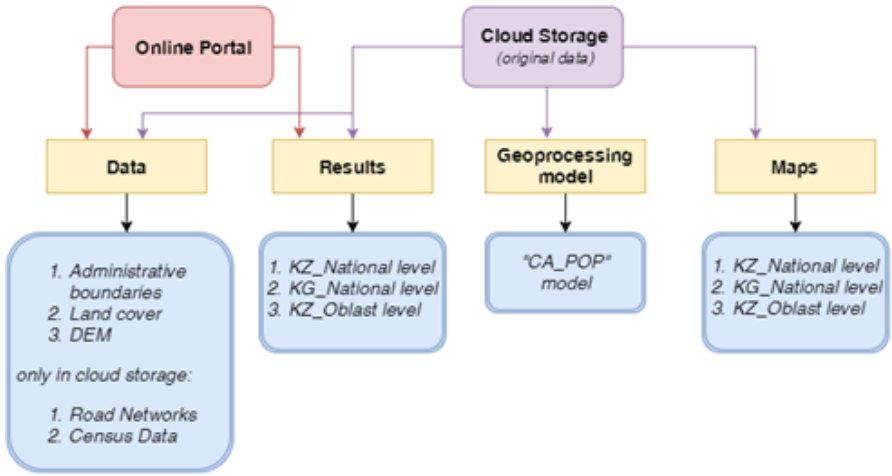


Figure 4: Structure of online portal and cloud storage

As mentioned before ancillary data are important for any population grid, but there is not possible to define which data are used for creating WPE dataset (particularly for Kazakhstan, or even for Central Asian region). Therefore, it is better to use data with defined sources. CA_POP model is based on the methods of CLC-Iterative and Street weighted approaches. Thus, there is double factor is taken into account for the determination of non-populated area, such as land cover and road networks. But also should be noted that DEM and Slope were also considered. Therefore, it is possible to say that chosen disaggregation approach gave accurate results than other existing sources and it is optimal for solutions of considered challenges. One of the main issues of this research is that sharing data and

results to Online Portal (ArcGIS Online) in order to do available for everyone. Because finding appropriate data and harmonizing it are time-consuming work. All datasets shared as Open Geospatial Consortium (OGC) standardized services. Since, in Online Portal raster data available only as tile layers (in this case), and since the created geoprocessing model was not shared as geoprocessing service, it is decided to save original data in Cloud Storage (for instance, Google drive) as well. Therefore also was decided to share data on Cloud Storage. In the end, Online Portal and Cloud Storage have following amount of data that is shown in Figure 4.

6. Conclusion

In the creating CA_POP disaggregation model was taken into account following criteria:

- that the model should be unified for all administrative units;
- the model should not be time consuming, therefore it split up for three sub-parts;
- the model does not only link to data, but also integrate it such as census data in excel format;
- the population percentage should be user-defined (if user has appropriate examined population percentage value);
- the model should be well-designed with corresponding metadata, therefore "Help" window was documented accordingly. Because any user should be able to use model.

The CA_POP model has some limitations together with pros. It should take into account that result of CA_POP model is not ideal, but it is comparable. Because for population percentage that defines how many people attributed to that or another class was taken from the paper by Gallego and Peedell (2001). These values of percentage are suited for European countries. Therefore, if there are necessities of more accurate data, then it will need to use population percentage values for selected study area (Kazakhstan). In addition, it was not to take into account the changes of roads by the census years. Since the Independence of Kazakhstan, it had been constructed 40 thousand kilometers of roads that somewhat may can influence to the settlements location. Therefore, in order to reach ideal results, there is need to use "ideal" input data.

Reference

Bierkens, M., Finke, P. and De Willigen, P., 2000, Upscaling and Downscaling Methods for Environmental research. Kluwer Academic Publishers, Dordrecht, The Netherlands.

California, C. F. S. S. A. T. U. O., 2016, *Database of Global Administrative Areas*. Version 2.8. Davis: Center for Spatial Sciences at the University of California.

Center for International Earth Science Information Network, C. C. U., 2016, *Gridded Population of the World, Version 4 (GPWv4): Population Density*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Deichmann, U., 1996, *A Review of Spatial Population Database Design and Modeling*. National Center for Geographic Information and Analysis.

Department of Geography, U. O. M., 1998, *UMD Global Land Cover Classification.* In: Hansen, M., R. Defries, J.R.G. Townshend and Sohlberg, R. (ed.) 1.0 ed. College Park, Maryland.

Eicher, C. L. and Brewer, C. A., 2001, Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science*, Vol. 28, 125-138.

Flowerdew, R. and Green, M., 1989, Statistical Methods for Inference between Incompatible Zonal Systems. *Accuracy of Spatial Databases*, 239-247.

Gallego, J. and Peedell, S., 2001, Using CORINE Land Cover to Map Population Density. Towards Agri-Environmental Indicators, *Topic Report*, Vol. 6, 92-103.

Gallego, F. J., 2010, A Population Density Grid of the European Union. *Population and Environment*, Vol. 31, 460-473.

Gallego, F. J., Batista, F., Rocha, C. and Mubareka, S., 2011, Disaggregating Population Density of the European Union with CORINE Land Cover. *International Journal of Geographical Information Science*, Vol. 25, 2051-2069.

Government of Canada, N. R. C., Centre for Topographic Information 2000, *Vector Map Level 0 (VMap0).* In: AGENCY, N. G.-I. (ed.).

Li, T., Pullar, D., Corcoran, J. and Stimson, R., 2007, A Comparison of Spatial Disaggregation Techniques as Applied to Population Estimation for South East Queensland (SEQ), Australia. *Applied GIS*, Vol. 3, 1-16.

Mayaux, P., Strahler, A., Eva, H., Herold, M. and Shefali, A., 2006, Validation of the Global Land Cover 2000 Map. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 44, 1728-1739.

Mennis, J., 2009, Dasymetric Mapping for Estimating Population in Small Areas. *Geography Compass*, Vol. 3, 727-745.

Mennis, J. and Hultgren, T., 2006, Intelligent Dasymetric Mapping and its Application to Areal Interpolation. *Cartography and Geographic Information Science*, Vol. 33, 179-194.

Miller, H. J. and Han, J., 2009, *Geographic Data Mining and Knowledge Discovery*, CRC Press.

Mrozinski, R. D. and Cromley, R. G., 1999, Singly- and Doubly-Constrained Methods of Areal Interpolation for Vector-based GIS. *Transactions in GIS*, Vol. 3, 285-301.

NGCC, N. G. C. O. C., 2010, GlobeLand30. Beijing, 100830, China.

Pérez-Hoyos, A., García-Haro, F. J. and San-Miguel-Ayanz, J., 2012, Conventional and Fuzzy Comparisons of Large Scale Land Cover Products: Application to CORINE, GLC2000, MODIS and GlobCover in Europe. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 74, 185-201.

Reibel, M. and Bufalino, M. E., 2005, Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environment and Planning A*, Vol. 37, 127-139.

Scholz, J., Andorfer, M. and Mittlboeck, M., 2013, *Spatial Accuracy Evaluation of Population Density Grid Disaggregations with Corine Landcover*. Geographic Information Science at the Heart of Europe. Springer.

Shu, Y. and Lam, N. S. N., 2011, Spatial Disaggregation of Carbon Dioxide Emissions from Road Traffic Based on Multiple Linear Regression Model. *Atmospheric Environment*, Vol. 45, 634-640.

Steinnocher, K., Köstl, M. and Weichselbaum, J., 2011, Grid-Based Population and Land Take Trend Indicators–New Approaches Introduced by the Geoland2 Core Information Service for Spatial Planning. 1-9, https://ec.europa.eu/eurostat/cros/system/files/S6P4.pdf_en.

US Environmental Protection Agency [Online], Available: https://www.epa.gov/enviroatlas/dasymetric-toolbox [Accessed].

UCL, E. A. B. T., 2010, Globcover 2009. In: UCLOUVAIN Team: Sophie Bontemps, P. D., Eric Van Bogaert and ESA Team: Olivier Arino, V. K., Jose Ramos Perez (Eds.). UCLouvain & ESA Team.

USGS, U. S. G. S. Shuttle Radar Topography Mission 1 Arc-Second Global: SRTM1N22W016V3. Sioux Falls, South Dakota: U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center.

Wu, S. S., Qiu, X. and Wang, L., 2005, Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience and Remote Sensing*, Vol. 42, 80-96.

Xie, Y., 1995, The Overlaid Network Algorithms for Areal Interpolation Problem. *Computers, Environment and Urban Systems*, Vol. 19, 287-306.