Spatiotemporal mapping of population in Europe

Filipe Batista e Silva[1], Konstantin Rosina, Sérgio Freire, Marcello Schiavina, Carlo Lavalle, Massimo Craglia

European Commission

Joint Research Centre

Via E. Fermi 2749

I-21027 Ispra, Italy

Extended abstract

1. Introduction and scope

Population is a crucial variable for the social and regional sciences, the geosciences and for policy support in many domains. Yet, our knowledge of its spatial distribution is still nowadays very incomplete. Population is a temporally dynamic variable, with major shifts in its distribution occurring in daily and seasonal cycles, resulting in rapidly changing densities.

Spatially detailed representations of residential population exist at EU level since several years. While these maps can be used as proxy for night-time population distribution, the distribution of population for other time frames is practically unknown at almost every spatial scale. Consequently, all applied sciences and policy support that require spatially detailed information on population distribution are based on only a fractional and static representation of reality. Overcoming this large knowledge gap is the main goal of an on-going research project at the European Commission Joint Research Centre – ENhancing ACTivity and population mapping (ENACT).

Ultimately, the ENACT project aims at developing and implementing a consistent and validated methodology to produce multi-temporal population distribution grids for Europe. Such datasets are expected to help expand the knowledge base of spatiotemporal population patterns across the continent. Multi-temporal population grids can indeed become a useful and straightforward (i.e. easy to integrate) input to regional science, predictive models in various fields (e.g. transport, land use, economic and environmental), as well as assessments related to the exposure to natural and man-made disasters. The final output of ENACT is a set of multi-temporal population grid maps that take into account the main seasonal and daily variations of population, consistent with the most recent census data (2011), and covering the EU28 countries. The target spatial resolution is 1 Km, which is sufficient for intra-regional and even intra-urban analysis.

The purpose of this communication is to present the state of the art as well as the main challenges and advances regarding spatiotemporal mapping of population distribution in Europe in the context of the ENACT project.

[1] Corresponding author. E-mail: filipe.batista@ec.europa.eu.

2. State-of-the art

The European Commission Joint Research Centre (JRC) has a long experience in population mapping and modelling. Since the early 1990's it has contributed to revolutionize the way population is represented and mapped at European level. From the early works of Gallego and Peedell (2001) to most recently Gallego et al. (2011) and Batista e Silva et al. (2013a), methods to map population have been experimented and refined, allowing the creation and update of maps representing residential population across Europe. Thanks to these early efforts, rudimentary – and to large extent ineffective – maps showing population density per administrative unit have been replaced by more realistic and useful depictions of population distribution at regular grid cell level, typically with a 100x100 meters resolution and usually referred to as 'population grids'. The data structure and high resolution of population grids allowed wider integration with other datasets in Geographical Information Systems (GIS), and thus have become indispensable datasets among both social and environmental researchers and spatial planners.

The main principle underlying the construction of such maps relies on the combination of two inputs: population counts usually available per administrative units or census zones, and a covariate of population distribution at higher spatial resolution, for example residential areas extracted from land use maps, building footprints, impervious surfaces, road network or even night-time lights from remote sensing imagery (for an extensive review of population estimation methods using GIS and Remote Sensing, see Wu et al. 2005). Using similar approaches, population estimates for the Urban Atlas polygons have also been created at the JRC at the request of DG REGIO (Batista e Silva et al. 2013b, Batista e Silva and Poelman 2016). In addition, the JRC has achieved to model future population distributions at European level under different scenarios using the LUISA territorial modelling platform, given demographic projections, within-region migration and local potential for urbanization (Batista e Silva et al. 2013c, Lavalle et al. 2016).

Despite the many improvements introduced by different researchers over the years, European population grids are – still nowadays – essentially static maps of 'residential population'. Residential population refers to the number of people who declare to reside in a given location. As such, when mapping residential population we are essentially mapping the distribution of population during the night time, assuming that most people stay in their declared places of residence during the night for shelter and rest, and excluding the fraction of people who work outside their residences during night time. Maps of residential population have developed quicker due to easier access to data: all European National Statistical Institutes (NSI), at least once every 10 years, count systematically the number of residents per census zone.

Residential population grid maps, although sufficient for a range of purposes, describe only a fraction of reality. The spatial distribution of people during the daytime or where people stay in different seasons is practically unknown for any spatial scale. Yet, such information is essential to an all range of applications (Martin et al. 2010). The location of population during the day is determined by the location of economic, social and leisure facilities which pull population off their residences, driving commuting flows and other forms of daily trips. Daytime population distribution thus varies greatly from night-time distribution. Contrary to night-time population – which, as already mentioned, can be straightforwardly inferred by official statistics on residential population – it is much more challenging to infer daytime population distribution.

Addressing the needs of emergency response, compatible day- and night-time population grids have been produced in the mid-2000s for the USA (McPherson et al. 2004; Bhaduri et al. 2007). In Europe, such datasets have been mostly lacking, with only a few countries systematically collecting base data and modelling population distribution on the daily cycle (e.g. Ahola et al. 2007). More recent research (Martin et al. 2015; Martin et al. 2010; Aubrecht et al. 2014; Smith et al. 2016) has been increasing the resolution of the temporal component and/or including a seasonal dimension for limited regional areas by mining conventional data. Other authors have explored the contribution of non-conventional data such as mobile phone activity records (Deville et al. 2014; Tatem et al. 2014) or 'geotweets' (Patel el al. 2016) for population mapping in selected countries, a task which is not without shortcomings and challenges. A relatively straightforward approach was proposed to estimate day- and night-time population distribution at high resolution for the cities in Urban Atlas (Freire et al. 2015); yet its quality depends largely on that of ancillary datasets and availability of local parameters, and its accuracy is still uncertain.

The challenges posed by spatiotemporal mapping and modelling of population distribution cannot be addressed effectively by conventional data sources alone (e.g. official statistics and reference land use datasets). Significant advances in this field can only be attained if data from conventional data sources are combined with data from emerging, non-conventional data sources in a coherent methodological framework. Non-conventional data sources may include volunteered geographical information (Goodchild 2007), web-based social networks (Aubrecht et al. 2016), thematic proprietary databases, mobile phone operator data, or even navigation systems.

Data mining from such (big) data sources is becoming a common task in many geospatial applications. Craglia and Granell (2014) reviewed a range of projects leveraging citizen science or crowd sourcing in the area of environmental monitoring and smart cities application. There are several projects that demonstrate the benefits of combining official and non-official sources. These include for example the GEO-Wiki project focusing on global land cover validated by local people in a game-inspired apps, and its urban application Cities Geo-wiki2 whose aim is to map the physical geography of all major cities in the world to link to weather and climate models. Another approach to augment local land use and building use databases with crowd-sourced information is documented by Spyratos et al. (2014) who used data from 'Foursquare' to monitor the dynamic changes in building use in commercial areas. Other studies have also shown that thematic geospatial layers can be obtained from disparate data sources and integrated with existing land use maps for improved detail (Batista e Silva et al. 2013d; Jiang et al. 2015). One of the most promising data sources to estimate population density comes from mobile network operators. Several studies have documented the importance and usefulness of this data source (see for example Steenbrugger et al. 2014, and Deville et al. 2014), but one key problem remains data access, which is normally negotiated with the data providers by individual researchers for specific projects.

Until a few years ago, population grids were almost unknown to most researchers even in domains with a strong spatial dimension. But population grids gained momentum very rapidly, and have become mainstream input for many analyses. Eurostat, which was initially reluctant to publish population figures using non-conventional zoning systems, kicked-off the GEOSTAT project in 2010 to promote the production, dissemination and use of 'gridded' population among NSIs. But despite all the referred advances, Europe is still lacking a wall-to-wall, spatiotemporal model of population distribution.

The mains challenges and associated with spatiotemporal population mapping can be summarized as follows:

(i) **Fast population dynamics**: People commute, travel, and migrate faster than ever before.

(ii) **Multifaceted concept**: During the night most of us are 'resident population'; during the day our (multiple) occupations are related to the (multiple) locations of probable presence.

(iii) **Data availability issues**: No official statistical sources exist. Daytime population needs to be inferred from multiple, indirect, and perhaps, new data sources.

(iv) **State-of-the-art**: Few case studies, often incomparable due to the use different methodologies and input data of different nature.

3. Data, methods and workflow

This research project aims at addressing the challenges that are required to obtain (1) the amount of people per type of activity, per time-frame (e.g. season/month), and per region, and (2) their likely location at a high spatial resolution. The first element requires regional data on residents, employees per sector of activity, students and tourists, which can be derived mostly from official sources. The second element requires spatial data on the location of activities (e.g. manufacturing, retail, health, education, leisure/tourism), available from various conventional (e.g. remote sensing imagery, land use maps) and unconventional data sources (e.g. volunteered geographical information, large proprietary geo-databases, and web-services). The final challenge is to validate the produced multi-temporal population distribution grid maps. ENACT's output will be compared against reference data from independent sources such as multi-temporal population grids derived from mobile phone operator data or other sources (e.g. produced by other researchers), and micro-data from National Statistical Institutes (NSIs) on location of employment.

The methodology is structured around 4 main pillars, each corresponding to a key challenge for the ultimate process of producing multi-temporal population grids for Europe:

(i) Estimate of the number of people that are likely to be inside or outside their residences during the day in each region;

(ii) Determine of the seasonal variation of total present population in each region;

(iii) Determine the locations of probable presence of population, and perform the population allocation;

(iv) Validate the quality of the resulting multi-temporal population grid maps.

The first challenge will be tackled by estimating total population per main activity sub-groups per NUTS3 region. We will therefore assemble the following official data from Eurostat: no. of workers per sector of economic activity, no. students per main educational levels, no. of tourists, and number of inactive people (and thus likely to be at home or in its vicinity during the day).

The seasonal variation of total present population in a region is assumed to be linked primarily to touristic flows. Therefore, inbound and outbound tourist flows of each region per season (or per month) must be taken into account. This will be addressed by first downscaling total yearly nights-spent per NUTS2 to NUTS3 using the no. of beds per NUTS3. Then, breaking down the NUTS3 yearly nights-spent per month using, as proxy, monthly flight traffic information from Eurostat. Monthly nights-spent per region can then be easily converted to average no. of tourists present in a given month by dividing total nights-spent in a month by the total no. of days comprised in that month. Finally, the number of tourists present in given a region in a given month are removed from their likely regions of origin.

Determining the locations of probable presence of population is a major challenge. In fact, there is no one-stop shop for such multi-sector, highly detailed locations of activities. ENACT needs to resort to various sources, proprietary and open-source, conventional and non-conventional. Specific population sub-groups (i.e.

employed persons per sector, students, tourists, and inactive) then need to be allocated to locations of probable presence. Data will be extracted from multiple sources, namely conventional (e.g. LULC maps, remote sensing), volunteered geographic information (e.g. OpenStreetMap), web-based services, and proprietary sources (e.g. TomTom, PLATTS, GISCO-EuroGeographics, etc.). At night-time, residents will be located to residential areas, and tourists to touristic accommodation facilities. At daytime, the various population sub-groups will be allocated to the relevant activity types based on a probability matrix.

The last (but not least) challenge of ENACT concerns the validation of its outputs. This will be done by comparing ENACT's outputs with other reference data, in particular emerging population density grid maps derived from mobile phone operator data, but also to existing employment grids produced by some NSIs.

Figure 1 summarizes the overall project workflow.

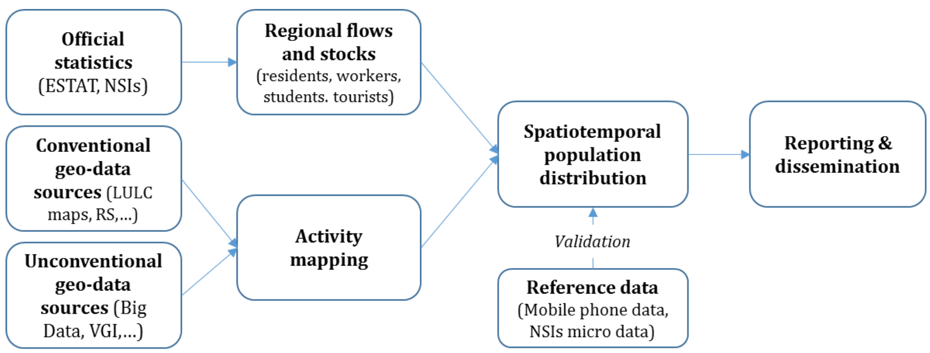

Figure 1. Overall project workflow.

4. Early conclusions and way forward

Although the project is still on-going, some preliminary conclusions can be drawn already. Multi-temporal modelling of population distribution is an exceptionally data-intensive task. Recently, the minimum required bulk of data to progress the task has been collected. Currently groundwork for the modelling is being laid by gap filling, data preparation and refinement. New data-related challenges are expected to appear as further datasets are explored.

Consequently, information drawn from numerous diverse data sources will be combined to fulfill the main goal of ENACT. Unfortunately, the inaccuracies present in each dataset will propagate and accumulate in the final product too. Knowing the error of the produced dataset is necessary to inform the users of the product but also to determine spatial scale at which it should be used. Therefore, designing a robust validation and quality assessment strategy will be no less important and challenging as the modelling per se.

As it is likely that the quality and quantity of suitable input data will grow in future, it is also vital to attempt at developing a flexible modelling framework that could accommodate gradual data (and methodological) improvements.

References

Aubrecht C, Özceylan Aubrecht D, Ungar JK et al. (2016). VGDI - Advancing the Concept: Volunteered Geo-Dynamic Information and its Benefits for Population Dynamics Modeling. Transactions in GIS.

Batista e Silva F and Poelman H (2016) Mapping population density in Functional Urban Areas - A method to downscale population statistics to Urban Atlas polygons. JRC Technical Report no. EUR 28194 EN. doi:10.2791/06831.

Batista e Silva F, Gallego J, Lavalle C (2013a) A high-resolution population grid map for Europe. Journal of Maps 9(1): 16-28.

Batista e Silva F, Lavalle C, Jacobs-Crisioni C et al. (2013c) Direct and indirect land use impacts of the EU cohesion policy. Assessment with the Land Use Modelling Platform. JRC scientific and policy report 26460. Luxembourg: Publications office of the European Union.

Batista e Silva F, Lavalle C, Koomen C (2013d) A procedure to obtain a refined European land use/cover map. Journal of Land Use Science 8(3): 255-283.

Batista e Silva F, Poelman H, Martens V, Lavalle C (2013b) Population Estimation for the Urban Atlas Polygons. JRC Technical Report 26437. Luxembourg: Publications office of the European Union.

Bhaduri B, Bright E, Coleman P and Urban M (2007) LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69: 103-117.

Craglia M and Granell C (2014) Citizen Science and Smart Cities. JRC Technical Report 26652. Publications Office: Luxembourg.

Deville P, Linard C, Martin S et al. (2014) Dynamic population mapping using mobile phone data. Proceedings National Academies of Science 111(45): 15888-15893.

Freire S, Florczyk A and Ferri A (2015) Modeling Day- and Nighttime Population Exposure at High Resolution: Application to Volcanic Risk Assessment in Campi Flegrei. In Palen, Büscher, Comes and Hughes (eds.) Proceedings of the ISCRAM 2015 Conference, Kristiansand, May 24-27 2015.

Gallego FJ and Peedell S (2001) Using Corine Land Cover to map population density. In European Environment Agency. "Towards agri-environmental indicators: Integrating statistical and administrative data with land cover information". Copenhagen. pp. 94–105.

Gallego FJ, 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 25(12): 2051–2069.

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4): 211–221.

Jiang S, Alves A, Rodrigues F et al. (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. Computers, Environment and Urban Systems. In press.

Lavalle C, Batista e Silva F, Baranzelli C et al. (2016) Land Use and Scenario Modeling for Integrated Sustainability Assessment, in J Feranec et al. (eds.) European Landscape Dynamics: CORINE Land Cover Data. CRC Press - Taylor & Francis Group. pp. 237-262.

Martin D, Cockings S and Leung S (2010) Progress report: 24-hour gridded population models. Proceedings of European Forum for Geostatistics 2010, Tallinn, Estonia.

Martin D, Cockings S and Leung S (2015). Developing a Flexible Framework for Spatiotemporal Population Modeling. Annals of the Association of American Geographers, 105(4), 754–772.

McPherson TN and Brown MJ (2004) Estimating daytime and nighttime population distributions in US cities for emergency response activities. In Symposium on Planning, Nowcasting and Forecasting in the Urban Zone, 84th AMS Annual Meeting, AMS, Seattle, 2004.

Patel NN, Stevens FR, Huang Z et al. (2016). Improving Large Area Population Mapping Using Geotweet Densities. Transactions in GIS.

Smith A, Martin D and Cockings S (2016) Spatio-Temporal Population Modelling for Enhanced Assessment of Urban Exposure to Flood Risk. Applied Spatial Analysis and Policy 9(2): 145-163.

Spyratos S, Lutz M and Pantisano F (2014) Characteristics of Citizen-contributed Geographic Information, in Huerta, Schade, Granell (eds.) Connecting a Digital Europe through Location and Place. Proceedings of the AGILE 2014 International Conference on Geographic Information Science, Castellon, Spain.

Steenbrugger J, Tranos E and Nijkamp P (2014). Data from Mobile phone operators: A tool for smarter cities? Telecommunications Policy 39(3-4): 335-346.

Tatem AJ, Huang Z, Narib C et al. (2014) Integrating rapid risk mapping and mobile phone call 545 record data for strategic malaria elimination planning. Malaria journal 13(52).

Wu SS, Qiu X and Wang L (2005) Population estimation methods in GIS and remote sensing: A review. GIScience and Remote Sensing 42(1): 80–96.