

## **Terra Communis (tComm): A free data provider for historical census micro-data.**

*A.M. Rodrigues<sup>(1)</sup>, B. Neves<sup>(1)</sup>, C. Rebelo<sup>(1)</sup>*

<sup>(1)</sup> e-GEO Research Centre for Geography and Regional Planning, Faculdade de Ciências Sociais e Humanas FCSH, Universidade Nova de Lisboa, Avenida de Berna 26-C, P 1069-061, Lisboa, amrodrigues, brunomaneves, crebelo @fcsH.unl.pt

### **ABSTRACT**

*The growing availability of census micro-data - demographic data aggregated for small-areas, allows detailed analysis of the social structure of small neighbourhoods. Such exercises are in most cases static since census tracts geometries change between every census exercise.*

*Using dasymmetric mapping techniques implemented within a spatially enabled PostgreSQL database, historical datasets were built for Portuguese census information, for the years 2001 and 2011. Information is freely made available using the Open Geospatial Consortium (OGC) Web Feature Service (WFS) Interface Standard. Geoserver is used to share data over the web and an interface is implemented using the JavaScript library OpenLayers.*

*The innovative nature of tComm is three-folded: (1) for the first time, a large scale comprehensive time-space database is built for Portuguese census data; (2) the fact that data is made available online contributes to the goal of making knowledge symmetrically available; (3) the fact the only Free and Open Source Software (FOSS) is used means that, other than man-hours, the project is costs' free.*

**Key words:** *Dasymmetric mapping, PostgreSQL, PostGIS, GeoServer, OpenLayers.*

## INTRODUCTION

Research breeds innovation, which in turn spills over society and space through contagious behaviour. Moreover, the importance of collaborative work within and between groups is justified as it promotes synergies. If on one hand this traditionally happens within a closed system because of agents' proximity, on the other, linkages into the civil society (namely into the administrative, productive and informal sectors) are non-existent if channels are not opened between producers of new (open) knowledge and the outside.

The locus - or focal point, of a given action, if isolated, brings no value-added into the surroundings communities. On the other hand, if channels are opened for participation of external actors, actions have a prolonged effect, which naturally depends on their significance. If the focus is on the activities of a research centre, then it is fair to say that with no channels to the outside, the end-product of research projects is of no use. Moreover, if communication between pairs is not symmetrical, research endeavours end up being redundant.

Direct applicability varies between pure science and directly applicable science. One is not more valid than the other, and the degree of needed exposition to the civil society depends. But if society invests scarce resources into scientific research, there must be some linkage effect between the production of scientific knowledge and society in general. Yet, there is one variable which should be always accounted for: creativity. Knowledge production is an act of creativity, and when scientific research is bounded by specific goals, creativity is hampered. Hence, scientific knowledge production lives in and between the objective/subjective realms.

Summing up, as with any human action, findings by researchers must be challenged by peers in order to become valid. Methods should be clear and freely available for replication. As should be results. Spillover effects into society are greater, the greater is the availability of knowledge.

In Human Geography, at the micro level (small-area data), historical census datasets have enormous utility. Applications range from areas such as geo-demographics, geo-marketing, insurances and local land planning (United Nations 2011). Even if the disaggregation level of analysis is lower - the level of aggregation is higher (for example local authorities or municipalities) - census tract data allows the analysis of the intra-variability for each of these regions, minimising the problem of ecological fallacy. In such cases, these micro-data serves as the structural analysis backbone of spatial units; and structure can be represented by a probability density function representing intra-regional variability.

When the shape of spatial units from distinct datasets does not coincide (Santos 2007), dasymmetric mapping using control units is necessary to build coherent information sets (Goodchild et al. 1993, Mennis 2003, Rodrigues et al. 2012a, Rodrigues et al. 2012b). If the starting point is data aggregated for different geometries and the objective is to build a coherent spatial-temporal database, some method must be used to overcome the issue of asymmetry in terms of spatial form. Using control units/zones, it becomes possible to create common geometries where data is re-

allocated according to these common shapes (Goodchild et al. 1993). The definition and use of control zones is of the utmost importance since they are the source of the weighting scheme used to re-allocated data for a common geometric set of spatial units. Also, the shape of the resulting areas is important since neighborhood effects depend on the contact-area between regions. Figure 1 shows an example non-coincidence in census geometries using Portuguese official data corresponding to three census exercises.



Figure 1: BGRI (“Base Geográfica de Referência de Informação”), BGRE (“Base Geográfica de Referência Espacial “)

The Terra Communis initiative uses exclusively Free and Open Source Software (FOSS) also named as Free/Libre Open Source Software (FLOSS) technologies. The word “Free” in FOSS does not mean that it is free of cost, but is referring to the software freedoms that are addressed to FOSS software (Steiniger and Hay 2009). These are stated in the Free Software Foundation (FSF) Website (<http://www.fsf.org>) and GNU Operating System Website (<http://www.gnu.org>) and refer to:

- (0) The freedom to run the program, for any purpose.
- (1) The freedom to study how the program works, and change it so it does your computing as you wish. Access to the source code is a precondition for this.
- (2) The freedom to redistribute copies so you can help your neighbour.
- (3) The freedom to distribute copies of your modified versions to others. By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

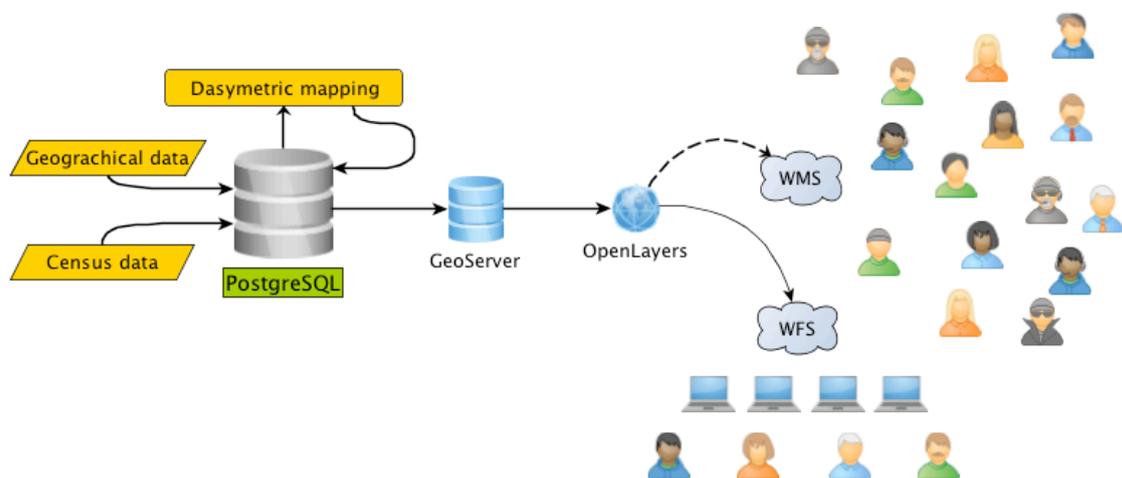


Figure 2

Project tComm's objectives are two-folded: first, it is intended to produce for the first time comprehensive datasets of historical datasets at the census tract level. Second, by providing these information freely over the World Wide Web, a significant contribution is being made in terms of the richness of demographic information available. This in turn is expected to have important spillovers in various areas of knowledge. A summarized workflow is shown in figure 2. One aspect which is a central motivation for the project is the purpose of closing the gap between academic research and civil society. Using exclusively Free and Open Source Software guarantees that transmission of technical know-how is symmetric.

## METHODOLOGY

Form, structure and function (Santos 1997) are fundamental aspects of spatial units' characterization which define how successful is the exercise of building multi-temporal micro data from census tracts' information collected for different timestamps. Broadening the level of conceptual analysis, form refers to "a particular way in which a thing exists or appears" (Abate and Jewell 2001); this implies attributes as shape, size, etc. The absolute or relative area occupied by a particular agent - or object, which serves as a control zone (ex: building, dwellings' frequency) is one key element of that agent. The structure (ex. distribution) of agents/ objects within a common spatial geometry (the result of transforming an asymmetric map into a symmetric one) influences the weight of each object in the control zones' definition. Finally, the function of the object determines how that object should be taken into account; for example, a vacant/empty building may be understood as having no function (commercial, residential, etc.). Figure 3 shows a snapshot of buildings' footprint which was used in this study as control zones.

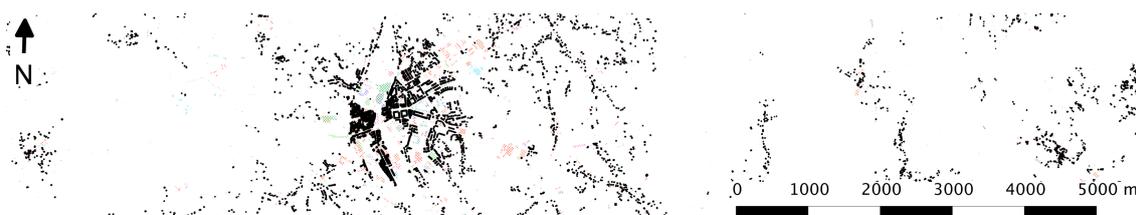


Figure 3

In order to illustrate the difficulties arising from distinct geometries, assume a simple spatial structure and a small number of agents/events (figure 4). The count dataset from both schemes represented by the point events on the left would result in the following counts:

$$Scheme1 = [8, 4, 2, 4]; Scheme2 = [5, 7, 6]$$

The quantification of inter-regional flows is also dramatically different as some flows are simply ignored as their origin and destination is contained within the same spatial unit.

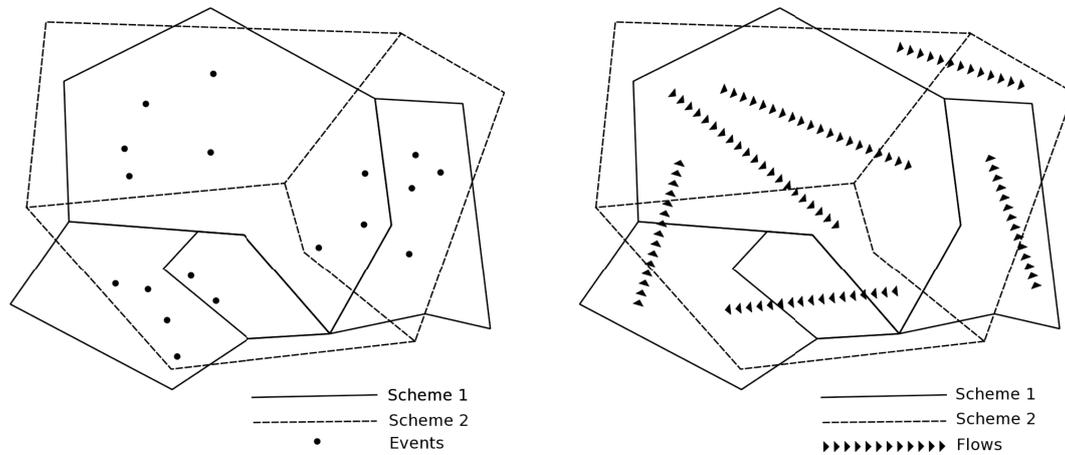


Figure 4

### Dasymmetric mapping

Any attempt to construct a multi-temporal census micro dataset is hampered by the non-coincidence of geometries. In order to transform one of these into a symmetric map, two steps must be taken: first, a common geometry must be adopted; second, data must be re-allocated according to this new geometry. Thus, there are two sets of geographical structures: the original geometry layer(s) and the target geometry.

Three related methodologies can be used, which differ in terms of complexity and are totally dependent on the availability of ancillary data: first, if it is assumed that distribution of agents/events over the original spatial units is uniform, then the re-allocation exercise results from weighting data according to the size (area) of each unit - *form* attributes are necessary and sufficient since data is assumed to be stationary over the spatial surface. Second, if we drop this assumption, then different weighting schemes must be designed. In general, this involves the use of control zones, representing detailed information on the distribution of some sort of objects which control and in fact determine the weighting scheme. Other than form, *structure* attributes are used; land- cover layers serve to discriminate between empty and non-empty spaces. Finally, if information is available in respect to land-use, stratification of objects according to their *function* becomes possible.

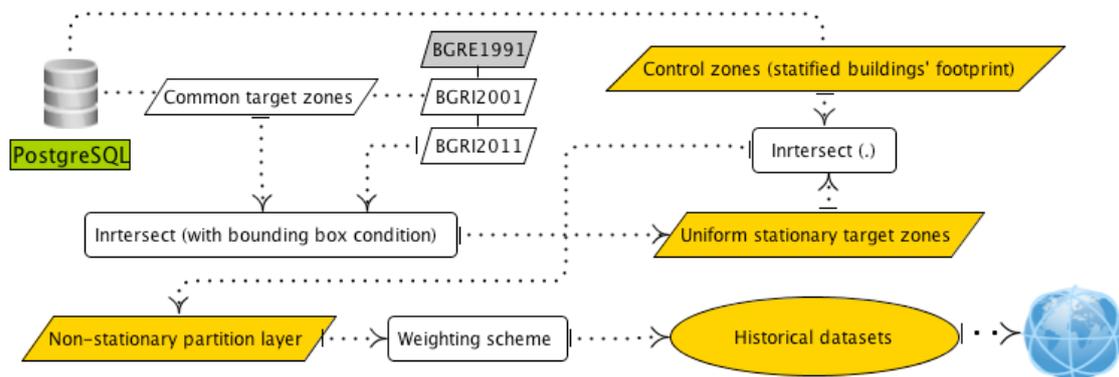


Figure 5

Figure 5 represents the different stages of the exercise. In the present work, the common target zones correspond to the BGRI 2001. Each unit's area shape the *form* attribute which would suffice if data was stationary. If building blocks were assumed to be of the same type (same function), then their area within each common target zone would represent the *structure* attribute. Building blocks for a case-study referring to the municipality of Tomar were stratified resulting in the resident buildings' footprint layer, which overlaid with the stationary target zones, resulted in non-stationary partition layer. This allowed the computation of proportions which allowed a proper weighting scheme to be designed and data to be re-allocated for the common target zones.

### Implementation

Using only FOSS, a framework was developed with the objective of creating a new channel for the diffusion of new geographical knowledge. Layers of information are integrated into a Relational Database Management system (DBMS). PostgreSQL with spatial capabilities implemented with PostGIS (<http://www.postgresql.org> and <http://postgis.refractory.net>), proved a competent platform for both data storing and analysis. Two Open Geospatial Consortium (OGC) standards allowed, using Geoserver (<http://www.geoserver.org>), the publication of information for viewing and downloading purposes (respectively WFS and WMS). Using an open-source javascript library, layers were integrated into a common system where information is accessible through any web-browser. Depending on the nature of particular information, the degree of interaction varied. Openlayers (<http://openlayers.org>) JavaScript library was used for creating a graphical interface.

### CONCLUDING REMARKS

This work demonstrated the strength of a spatially-enabled DBMS in terms of storing and of its spatial analysis capabilities. The SQL scripts used proved to be extremely efficient in dealing with large geometry tables for the implementation of dasymmetric mapping techniques.

Terra Communis (tComm) is an initiative strongly rooted on the belief that information should be made available to the general public meeting high quality

standards and though easy-to-use interfaces. The initiative became operational using only Free and open source software (FOSS). The obvious costs' reductions are perhaps the less important advantage of such strategy; FOSS guarantees accountability and promotes knowledge spillovers.

tComm is not an end in itself; information is only of any use if and when it is applied for the production of new knowledge and the understanding of a given reality.

## REFERENCES

FREE SOFTWARE FOUNDATION (n.d.), Free Software Foundation – what we do, In Free Software Foundation, Retrieved March 10, 2012, from <http://www.fsf.org/>

Goodchild, Michael F. (2007). Editorial: Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2.0. *International Journal*, 2(2), 24-32. doi:10.1016/j.jenvrad.2011.12.005

Goodchild, M.F., Anselin, L., Deichmann, U.: A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* 25(3), 383–397 (1993)

GNU OPERATING SYSTEM (2012), What is free software?, The Free Software Definition, In GNU Operating System, Retrieved March 10, 2012, from <http://www.gnu.org/philosophy/free-sw.html>

Mennis, J. (2003). Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1), 31-42. doi:10.1111/0033-0124.10042

Rodrigues, A., Santos, T., Deus, R. F. D., & Pimentel, D. (2012). Land-Use Dynamics at the Micro Level: Constructing and Analyzing Historical Datasets for the Portuguese Census Tracts. In B. Murgante, G. Borruso, & A. Lapucci (Eds.), *ICCSA Proceedings* (pp. 565-577). Salvador da Bahia: Springer-Verlag.

Rodrigues, A., Santos, T., & Pimentel, D. (2012). Asymmetrical-mapping based methodology : Constructing historical datasets for Portuguese census tracts. In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), (pp. 24-27). *Proceedings of the AGILE'2012 International Conference on Geographic Information Science*, Avignon, April, 24-27, 2012. doi:ISBN: 978-90-816960-0-5

Santos, A. (2007). UNITED NATIONS SECRETARIAT Department of Economic and Social Affairs Statistics Division May 2007 English only United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems United Nations , New, (May).

Santos, M.: *Espaço e Método*, 4th edn. Nobel (1997)

Steineger, S.; Hay, G.J.; (2009), "Free and open source geographic information tools for landscape ecology", *Ecological Informatics*, Volume 4, Issue 4, September 2009, Pages 183-195, ISSN 1574-9541, <http://dx.doi.org/10.1016/j.ecoinf.2009.07.004>

United Nations - Department of Economic and Social Affairs Statistics Division; Handbook on Geographic Databases and Census Mapping (Draft version). In: Proceedings of the United Nations Expert Group Meeting on Measuring the Economically Active Population in Censuses, New York (2008)

## **ACKNOWLEDGEMENTS**

This paper/article/chapter presents research results of the Strategic Project of e-GEO (PEst-OE/SADG/UI0161/2011) Research Centre for Geography and Regional Planning funded by the Portuguese State Budget through the Fundação para a Ciência e a Tecnologia. It was also funded by FCT grant SFRH/BDP/66012/2009).