SLkit: An R package for property extraction and analysis of multiple Sensing Layers*

Fabrício Ferreira[1], Thiago H. Silva[2], Antônio A. F. Loureiro[1]

[1]Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil

[2]Department of Informatics
Federal University of Technology - Paraná
Curitiba, Brazil

{fabricio.silva,loureiro}@dcc.ufmg.br, thiagoh@utfpr.edu.br

Abstract. *The popularisation of smartphones and the increased use of mobile applications allow users to become not only consumers but also data producers. Data shared voluntarily by users through mobile applications open unprecedented opportunities to generate useful knowledge, identify and solve issues, and provide new services. Due to the complexity of urban phenomena and the large volume of data available, stages such as modelling, preprocessing and analysis of sensing layers can be time-consuming. To help to tackle that issue, this study presents an R package intended to give support to researchers regarding decision making and evaluation of sensing layers. It provides functions for property extraction and multilayered analysis which can be customised according to one's project needs, helping to leverage new applications that explore the concept of sensing layers.*

1. Introduction

The number of people who live in urban areas outnumbers, since 2009, the portion of rural residents. This difference continues to increase, mainly in developing countries where people move to the city in search for a job opportunity and better quality of life [DESA 2015]. As a result government authorities have focused on projects that face challenges associated with urban sprawl regarding basic services such as energy, housing, transportation, waste management, water and sanitation [Khatoun and Zeadally 2016].

Another fact to be considered is the evolution of the use of mobile gadgets and how it is changing the way people relate to the digital and physical world. Smartphones are equipped with powerful sensors used in various applications. Some of these use geographic location service to improve user experience, allowing users to generate a large volume of data. This data can contain information capable of characterising personal preferences (e.g., places one enjoys going) as well as collective behaviour within a specific location (e.g., how people move around a neighbourhood). In this scenario, users are considered sensors themselves, composing a Participatory Sensor Network (PSN),

*The code and documentation are available on https://github.com/FdeFabricio/POC. The project is licensed under the GPL. Any contribution/comment is welcome.

which have a wider coverage and lower cost than traditional Wireless Sensor Networks (WSN) [Silva et al. 2014a]. Hence urban solutions which take into account such data have been shown to be more effective and more democratic.

The large volume of data regarding diverse aspects of a city represents an unprecedented opportunity for assimilating knowledge, identifying and solving problems, besides providing new services. In order to achieve that, sophisticated methods of data analysis are required, capable of dealing with the heterogeneity of data sources, a common challenge of this context. A possible approach, discussed in Section 2, is by modelling each dataset as a Sensing Layer (SL), which *"consists of data describing specific aspects of a geographical location"* [Silva et al. 2014b]. The majority of urban phenomena have substantial complexity and could not be described using a single layer. Urban mobility, for instance, can be influenced by the weather, topography, big events, distribution and level of public transportation service, income distribution, among others. Therefore, it is fundamental to adopt a multilayered approach in order to characterise complex phenomena and maybe try to predict them more accurately. However, the use of multiple layers face challenges pointed out by [Silva et al. 2015] such as ensuring spatiotemporal consistency between all related layers, compiling data by location and/or users and comparing data with distinct lifespan.

Pre-processing and analysis of sensing layers are important steps since both potentially valuable insights and prevent inappropriate use of sensing layers, such as spatiotemporal incompatibility and illusory correlations. This stage is often done manually or through specific scripts, which are not commonly reused seamlessly in other projects, becoming a repetitive and redundant effort. Thus, this work presents a tool capable of automating important processes of spatial and temporal properties extraction and analyses on multiple SLs. The tool consists of an R package and it is intended for researchers who wish to save time on projects involving a considerable number of spatiotemporal datasets.

This paper is organised as follows. Section 2 defines Sensing Layers, how they can be used and their challenges. Section 3 presents the related work. Section 4 discusses temporal and spatial property extractions on the package. Section 5 discusses multilayered analyses. Finally, Section 6 points out the conclusions and future work.

2. Sensing Layers and its challenges

A SL is a set of data related to the same domain, describing certain aspects of a particular geographic region [Silva et al. 2014b]. For instance, layers may contain information about weather, traffic condition, social networks, check-ins, etc. Such data can be obtained from a variety of sources, such as PSNs, Web services, or even conventional WSNs. In general, they are composed of parameters such as: timestamp in which data were measured or collected; geographical location it is inserted into; agent identification (e.g., sensors or user ID of a social network); elements describing domain specificities and the value of the measured variable.

The use of SLs allows continuous monitoring of an area over time on assorted aspects, as well as information retrieval that would not be possible to conduct by analysing each dataset separately. Such representation makes it easier to investigate his-

torical data and identify patterns. The application of individual sensing layers already shows value, as mentioned in Section 3. However, the use of multiple layers allows data contextualisation which consequently allows a better understanding of complex phenomena.

There are challenges in using multiple SLs though. The lack of spatial or temporal correspondence may lead to an illusory correlation between different layers. In order to conduct correct analysis it is fundamental that the datasets are in a compatible format and that there is spatiotemporal correspondence, i.e., the data ought to be from the same area and period of time. It is also possible that the data distribution is not uniform lacking data in specific time frames (late at night or bank holidays, for example) or in specific areas (e.g., industrial neighbourhood and suburbia), which makes it problematic to draw conclusions or to validate hypotheses. Typically spatial and/or temporal resolution, i.e., how frequent a data is updated and how many square meters it represent, are not equivalent which also can impact on the multilayered correlation. Depending on the context, different conditions such as the above mentioned must be satisfied to legitimise a multilayered study, which proves the importance of the property extraction and analysis stage.

3. Related work

It is possible to extract information when conducting data analysis on a single layer. However, the lack of complementary data can restrict the study case or even lead to significant error. For example in 2008 Google launched Google Flu Trends (GFT), a Web service to estimate the number of flu cases by using data from its search engine. The initial estimates provided overshot more than 50 % of the number of cases and even after adjustments the error was still greater than 30 % [Lazer et al. 2014]. On the other hand, the GFT worked highly accurate when used in conjunction with data from the Centers for Disease Control and Prevention (CDC).

[Silva et al. 2014b] present a study about using PSNs as SLs. Using two-layered examples, the authors discuss the process of data collection, modelling and operations for information retrieval. The first case uses data from Instagram and Foursquare to identify points of interest in the city of Belo Horizonte, Brazil. The second correlates sentiments found in comments on Foursquare about establishments in a particular area with the average income of its inhabitants, indicating the absence of negative feelings in areas with greater purchasing power.

[Bakhshi et al. 2014] analysed Yelp reviews, correlating them with weather, demographics and local characteristics. In this study, the authors highlighted factors which directly influence clients emotions, which also indirectly affect their evaluation.

[Machado et al. 2015] use multiple Sensing Layers to analyse urban mobility in six cities (New York, Chicago, Los Angeles, Paris, London and São Paulo). Based on weather data from the Weather Underground service in addition to a database of check-ins made in these cities, the authors attempt to find a relationship between climate and urban behaviour analysing approximately 120 days. The findings show that there is a phase transition phenomenon, supported by the correlation between temperature variation and mobility pattern in Paris and London. In other cities where such phenomenon

was not well structured, the variation of the movement pattern was still identified when considered small regions within the city.

In the scope of urban mobility, there are studies that analyse databases on traffic flow and its correlation with social networks. [Tostes et al. 2013] use traffic data from Bing Maps to predict traffic conditions, such as congestion, with data from the preceding week. [Ribeiro et al. 2014] try to understand if Foursquare and Instagram data can be used to understand traffic in cities. This work shows that there is a great correlation between behavioural data of social networks (social sensor) and traffic conditions, being really useful, again, for predictions.

Although there are R packages created for data science projects intended to pre-process geospatial data and even make map plots, there is none for sensing layer analysis. The package presented in this work relies on generic R packages such as dplyr, ggplot2, rgdal, rgeos and ggmap and provides functions to be used and customised, if needed, on multilayered data projects.

4. Property Extraction

Different types of data can be collected from distinct sources. Temperature, for example, can be obtained by distributed sensors in a city, meteorological stations, prediction by satellite image analysis, PSNs, among others. Using all these sources to represent the same variable might configure an unnecessary effort, which could impact directly on the study complexity. Therefore, it would be necessary to extract characteristics from each dataset and select the one, or those, that best meet the purpose of the work. Another important aspect is to evaluate the spatial and temporal coverage of the data and its degree of repeatability: you may have a large amount of data, but 20% of those may be sufficient to represent that variable reliably.

For the reasons given above, it proves necessary to quantify characteristics of SL to serve as a basis for decision-making, in addition to performing analyses, discussed in Section 5.

The term property, in this work, is any spatial or temporal characteristic of a Sensing Layer that can be measured and compared among different layers. It was not found in the literature a list of properties that can be extracted from a SL. Therefore this work presents a set of properties, divided into temporal and spatial, selected based on the methodology of the related work. Not that this list can also be expanded when necessary.

4.1. Temporal Properties

This set of properties is a mechanism to characterise how data behave considering time as a factor. This work implements the functions presented below. The documentation, parameter listing, output plots and examples of use are in the project repository[1].

- **Temporal Coverage (tpCoverage):** temporal interval the data is inserted into. This function receives a dataframe, sweeps all its data and returns a vector with the earliest and latest timestamp.

[1] https://github.com/FdeFabricio/POC

- **Temporal Distribution (tpDistribution):** how much the data is distributed through time. This function divides the temporal coverage according to an input resolution and returns the percentage of it which has data associated with. For example for a crime dataset, if the time resolution is set as "daily", the function divides the data into separate days and return the number of days that have at least one crime record over the total of days.
- **Refresh Rate (refreshRate):** how often the data is updated. This function returns the average of the time difference between consecutive measurements.
- **Temporal Popularity (tpPopularity):** how much data are in each time unit considering a given resolution as input (hour of the day, day of the week, etc.) The function returns a histogram with the percentage of total data of each time interval.

4.2. Spatial Properties

These properties quantify the relationship with the geographic space. This work implements the functions presented below. The documentation, parameter listing, output plots and examples of use are in the project repository[2].

- **Spatial Coverage:** area the data is inserted into. This function analyses the latitude and longitude values of an input dataframe and returns the extreme coordinates forming a bounding box. It can also return a plot with the data and the bounding box on a map.
- **Spatial Distribution:** how much the data is distributed through space. This function divides the spatial coverage into a number of rectangles (given as input) and returns the percentage of them that has data associated with. For example for the crime dataset with a spatial coverage of 100 km^2, if the spatial resolution is 50, this function divides the space into 2 km^2 rectangles and returns the number of rectangles that have at least one crime record inside it over the total of rectangles.
- **Spatial Popularity:** this function returns a heat map, highlighting areas that have more data (high popularity) and empty ones (low popularity).

5. Multilayered Analysis

After analysing each dataset individually it is important to see how the selected layers relate to each other. A multilayered analysis intends to attest if the variable represented by one layer have any effect on another layer. Taking as an example two datasets of bicycle rental and precipitation level, one can try to verify if the volume of rain influences the use of such transport mode. To do so it is necessary to calculate the correlation of the two variables (rainfall in mm and number of rents, for example), which can be negative (the more it rains, the fewer people rent bikes), positive (the more it rains, the more people rent bikes) or no correlation (the rain does not affect the number of bicycle rentals).

To run a multilayered correlation, one must make sure that there is no incongruity between layers. By adopting data on the number of rented bicycles during the month of January and the rainfall index of August, one makes the mistake of trying to

[2]https://github.com/FdeFabricio/POC

```
$temporal
             checkin  instagram weatherUn noiseTube
checkin   1.00000000 0.99992646         1         0
instagram 0.99999316 1.00000000         1         0
weatherUn 0.03718432 0.03718184         1         0
noiseTube 0.00000000 0.00000000         0         1

$spatial
            checkin instagram noiseTube
checkin   1.0000000 0.9965368         0
instagram 0.9989773 1.0000000         0
noiseTube 0.0000000 0.0000000         1
```

Figure 1. Example of a four-layered STIA output

relate data that is time-separated and therefore could not have any causal relationship. The same mistake would occur if the layers correlated were separated in space (e.g., London bicycle rental data and Singapore's rainfall index).

To ensure that scenarios similar to those mentioned above do not occur, it is necessary to extract temporal and spatial properties of sensing layers and compare them. The package provides a SpatioTemporal Intersection Analysis (STIA), which aims to measure the level of intersection between different layers regarding spatial and temporal data. The function receives the different layers as parameters and returns a intersection matrix with the percentage of intersection between each layer. The result can be used to trim a layer when there is any incompatibility or even for decision making (to test if a certain dataset is suitable for the actual study).

Figure 1 presents an output of a STIA of four layers: Foursquare check-in data, georeferenced Instagram posts, precipitation by Weather Underground and noise level by Noise Tube. The resulting matrix shows that: 1) weatherUn has no geospatial data; 2) checkin and instagram layers are from the same area and period of time (spatiotemporal intersected); 3) the time period of checkin and instagram layers represents 3.7% of weatherUn's complete time interval; 4) noiseTube layer has no spatial nor temporal intersection with any other layer. Therefore, for a study with such layers, one would use only part of the weatherUn layer and none of the noiseTube.

Other analyses can be conducted considering the properties extracted. For instance, by extracting the seasonality and the temporal coverage of a layer, it's possible to identify if there are redundant data. Temporal and spatial popularity can be used to understand how users behave in a given social network. [Silva et al. 2013] identified, for instance, that Instagram data presented typical peaks at mealtimes, which is a valuable information when considering associating this data with other sources. Spatial popularity can be useful in a congestion study since PSNs such as Waze do not have

much data in the periphery nor small-sized cities. This means that any prediction will take into account only evidence from the city centre and main streets, which can be problematic.

Although the quality of a data source is used to decided if a layer will be considered in a project or not, it is generally difficult to determine an approach to measure it since it depends on the idiosyncrasy and the aims of a particular study. For example, a researcher can make a decision based on refresh rate when the variable measured constantly changes (e.g., bus position). Or wishing to create a robust predictive traffic model, one would choose layers with higher spatial and temporal distribution which, in other words, have data disperse on more parts of a the city (e.g., city centre and residential areas) and in different situations (e.g., peak time, night time, weekends, holidays). It is also possible that one would choose a layer with fewer data to minimise computational complexity. For example, in a study where no more than the average daily temperature is required, there is no need for a layer with temperature measurements every 5 minutes.

Different analyses can be conducted for each research and it depends entirely on the scope, selected layers and study goals. This package was thought to be incremented as necessary and free to any contribution. Since it is hosted on GitHub, one can download this package, import it to its project, make alterations to adapt the functions to the context of their study and even share such modifications back with the community. In addition, this package not only helps on the process of sensing layer research but also promotes collaboration between academics in an open source platform.

6. Final Remarks

This paper described a package developed in R to give support on the stages of modelling, pre-processing and analysis of sensing layers. It presents a set of functions capable of performing spatial and temporal property extraction and multilayered analysis intended to minimise effort on such studies dealing with sensing layers. The documentation of the project presents an installation tutorial, description of the functions and examples of their use. The package may need adjustments to adapt to the context and objectives of a given project. For this reason, an open source package becomes suitable since different researchers can make changes to the code and contribute to its improvement.

For future work it would be interesting to evaluate the level of acceptance of this package and if users are benefiting from it. Licensing the code with the GNU General Public License and providing the project on a collaborative development platform such as GitHub is a good alternative to engage collaborators and receive feedback.

Another possibility of future work includes the study of scenarios other than urban phenomena in order to attest the necessity of other property extractions. The package may need adjustments to deal with layers that have no geographical or temporal data since currently those are eliminated from the methods implemented.

References

Bakhshi, S., Kanuparthy, P., and Gilbert, E. (2014). Demographics, weather and online reviews: A study of restaurant recommendations. In *Proceedings of the 23rd Inter-*

national Conference on World Wide Web, WWW '14, pages 443–454, New York. ACM.

DESA (2015). *World Urbanization Prospects: The 2014 Revision*. United Nations, Department of Economic and Social Affairs, Population Division, New York.

Khatoun, R. and Zeadally, S. (2016). Smart cities: Concepts, architectures, research opportunities. 59(8):46–57.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, 343(14 March):1203–1205.

Machado, K., Silva, T. H., de Melo, P. O. V., Cerqueira, E., and Loureiro, A. A. (2015). Urban mobility sensing analysis through a layered sensing approach. In *2015 IEEE International Conference on Mobile Services*, pages 306–312. IEEE.

Ribeiro, A. I. J. a. T., Silva, T. H., Duarte-Figueiredo, F., and Loureiro, A. A. (2014). Studying traffic conditions by analyzing foursquare and instagram data. In *Proceedings of the 11th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, pages 17–24, New York. ACM.

Silva, T., Vaz De Melo, P., Almeida, J., and Loureiro, A. (2014a). Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, 21(1):42–51.

Silva, T. H., de Melo, P. O. V., Almeida, J. M., Viana, A. C., Salles, J., and Loureiro, A. A. (2014b). Definição, modelagem e aplicações de camadas de sensoriamento participativo. In *Proc. of XXXIII SBRC'14*, Florianópolis, SC.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Borges Neto, J., Tostes, A. I. J., Celes, C. S. F. S., Mota, V. F. S., Cunha, F. D., Ferreira, A. P. G., Machado, K. L. S., and Loureiro, A. A. F. (2015). Redes de sensoriamento participativo: Desafios e oportunidades. In *Magnos Martinello; Moises Renato Nunes Robeiro; Antônio Augusto Aragão Rocha. (Org.). Minicursos / XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 266–315, Porto Alegre. SBC.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013). Uma Fotografia do Instagram: Caracterização e Aplicação. In *Proc. of XXXII SBRC'13*, Brasília, DF.

Tostes, A. I. J., de LP Duarte-Figueiredo, F., Assunção, R., Salles, J., and Loureiro, A. A. (2013). From data to knowledge: city-wide traffic flows analysis and prediction using bing maps. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 12. ACM.