D4D Challenge
Commuting Dynamics 4 Change

R. Lario[*] M. Muñoz[†] R. Abad [‡] J. Gonzalez [§] A. Martín[¶] R. Maestre [‖]

Paradigma Labs Research Group

Paradigma Tecnológico

E. Perez [**] I. del Bosque [††]

Geographic Information Systems Unit

Spanish National Research Council

4-Jan-2013

Abstract

Our idea is to use the geolocation data from the antennas processing the mobile phones calls in order to know which sub-prefectures the customers have been getting around. The main goal of our project is developing spatio-temporal models to detect commuting patterns for the different sub-prefectures, including some other factors related to the region and/or time: wealth, development, infrastructure, investment, grants...

By means of GIS technology, we will be able to apply our generated models to the gathered data and to analyze their correlations over the Côte d'Ivoire surface, working with geographical layers: landcover, roads map, railways lines, water sources... Consequently, the reached conclusions from our study will be properly visualized, allowing a better explanation of the facts. With a bigger amount of data gathered for a longer period, more interesting and accurate trends could be discovered, allowing us to calculate associated coefficients.

Our analysis models will provide coherent data to support a correct urban design and will mean a monitoring tool for development, specially related to population dynamics. In the near future, some other measures could be included. For instance, hospitals and police stations locations, their calls rate... Thus, we could know its real use, being able to improve their service to the citizens: dangerous areas, crowded hospitals...

[*] rlario@paradigmatecnologico.com

[†] mmunoz@paradigmatecnologico.com

[‡] rabad@paradigmatecnologico.com

[§] jgonzalez@paradigmatecnologico.com

[¶] amartin@paradigmatecnologico.com

[‖] rmaestre@paradigmatecnologico.com

[**] eperez@cchs.csic.es

[††] idelbosque@cchs.csic.es

WHY FACE THIS CHALLENGE ?

Paradigma Labs Research Group[1] (PLRG) core values are conducted by one motivation: "To figure out the fuzzy dynamics between Humanity and Technology", providing tools and methods to study, display and understand these dynamics. Therefore, an international challenge whose research subject can be chosen freely as long as it relates to an objective of development improving quality of life for people, quickly held our attention.

Geographical Information Systems unit[2] at Spanish National Research Council[3] (GIS-CSIC) is a multidisciplinary group with a huge experience in Remote Sensing and Geoprocessing, providing a quality support for plenty of researches carried out at CSIC. Our team has been working hard to achieve the goal proposed in this paper.

Our final aim has been detecting spatio-temporal patterns in order to obtain an useful knowledge to better manage the country resources. For example, if we could predict the traffic intensity segmented by road, week day and hour, then another secondary roads could be suggested or the budget for the most used ones could be increased. Through mobile communications, a specific user can be tracked along the day, not only by means of 'call comunications' but also thanks to applications running on their handsets: IMS, RSS... The dataset provided by Orange is a sample, and it only uses 'call communications', however, with the whole set of data (i.e.: app and call comunications), we strongly believe more accurate and complete models could be discovered helping new kind of dynamics to be identified.

From our own experience studying and modeling several kinds of Human Dynamics like during the ESF project DynCoopNet(Solana and Alonso, 2012) and while developing a Business Intelligence Tracking Tool on Twitter (Marin et al., 2012), we can claim there are two main exploring perspectives: the Geographical one and Temporal one. We believe a mathematical model related to Human Dynamics must be managed with these two viewpoints. The Temporal component is useful by providing a tool to go backward and forward in order to get a more detailed understanding of the dynamic, not only moving across the timeline, but also creating temporal windows to group events. The Geographical component provides a more high-level understanding related to the human mobility across the space in different levels and relating it to some other spatial features. Mixing both components in a final and single visualization has led our study during the project.

Consequently, in this paper, we propose a Space-Temporal Model. A Space Model, because user interactions with geolocated antennas are analyzed and treated, and a Temporal Model since several time windows are used to group these user dynamics. The combination of these two variables is used and displayed by a Geographic Information System, GIS. Initially, several results are showed supporting the project main conclusions. However, what's really important is the whole process for handling the data, that is, the code, tools and methodology, which will be available to the researcher comunity, allowing to study more deeply the dynamics. For instance, a Standard Kernel Density estimation (KDE) aims to produce a smooth density surface of spatial point events over a 2-D geographic space(Bithell, 1990; Alegria et al., 2011), final dynamics visualization across the several days of the week will show by means of KDE, in order to understand and proof which an where are the maximum commuting peaks.

We have focused on the Commuting concept, which could be defined as follows: *Commuting* is regular

[1] http://labs.paradigmatecnologico.com/
[2] http://humanidades.cchs.csic.es/cchs/sig/
[3] http://www.csic.es/

travel between one's place of residence and place of work or full-time study (Wikipedia, 2012), but sometimes it refers to *any regular or often repeated traveling between locations when not work related.* Our first commuting approach is defined like: "Mobility patterns through inferring dynamic users movements grouped by temporal windows".

A 'dynamic user' is defined as an user changing his antenna location within the studied temporal window (i.e.: each temporal window groups the whole user communication during one specific hour). Among these temporal windows, 'static users' have been removed, i.e.: users who do not change their antennas locations within the temporal range. The justification to remove these users comes to focus our study on users that are moving into this temporal windows and perform micro-displacements. It is common that a unique user performs these two kind of dynamics for a same temporal window.

Mathematical model

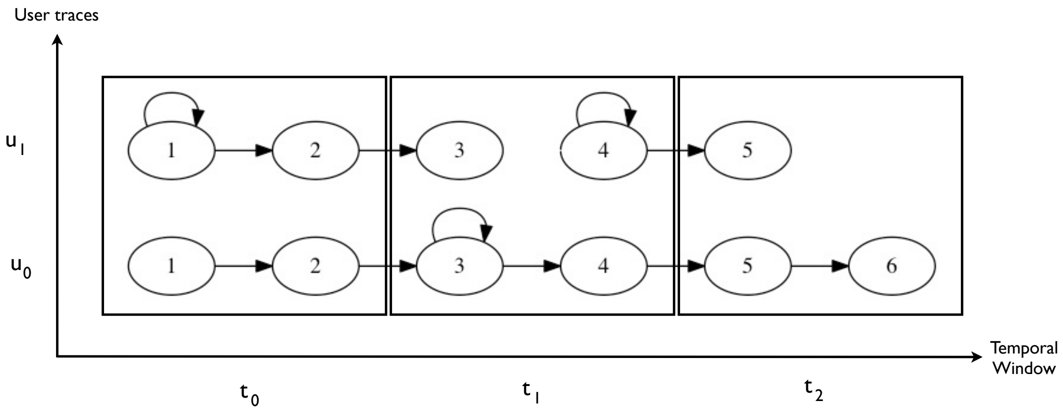


Figure 1: Dynamic and static user patterns

The above figure shows two users u_0, u_1 repesented on the vertical axe, grouped by three time windows t_o, t_1, t_2. A time window is defined as t_n where $n \in \{0, .., 24\}$. Each t_n groups the communications traces of the whole set of users in a 60 minutes range.

Formally, an user trace during a particular time window is defined as follows:

$$\vec{T}_{ut} = (p_0, p_1, ..., p_n)$$

where, $p_n \in \mathbb{R}^2$, $n \geq 2$, t is a temporal window range and, u the unique user id.
For instance, $\vec{T}_{00} = (p_1, p_2)$, $\vec{T}_{10} = (p_1, p_1, p_2)$ and so on.

Also, two functions are defined in order to measure the distance (Cook, 2012), given a set of points in spherical coordinates (i.e.: user$_{ut}$):

$$D(p_0, p_1) = \text{acos}(\sin(\phi(p_0^0)) * \sin(\phi(p_1^0)) * \cos(\theta(p_0^1) - \theta(p_1^1)) + \cos(\phi(p_0^0)) * \cos(\phi(p_1^0)))$$

where:

$$\phi(x) = (90 - x) * \frac{\pi}{180}$$

$$\theta(x) = x * \frac{\pi}{180}$$

therefore the function related to the distance is defined as follow [result in Km]:

$$U(u, t) = 6373 * \sum_{i=0}^{n-1} D(\vec{T}_{ut}^{\,i}, \vec{T}_{ut}^{\,i+1})$$

The second function is related to the number of antenna connections into a trace, the key point is to count only the dynamic transitions, i.e.: remove the self edges over a given trace as follows:

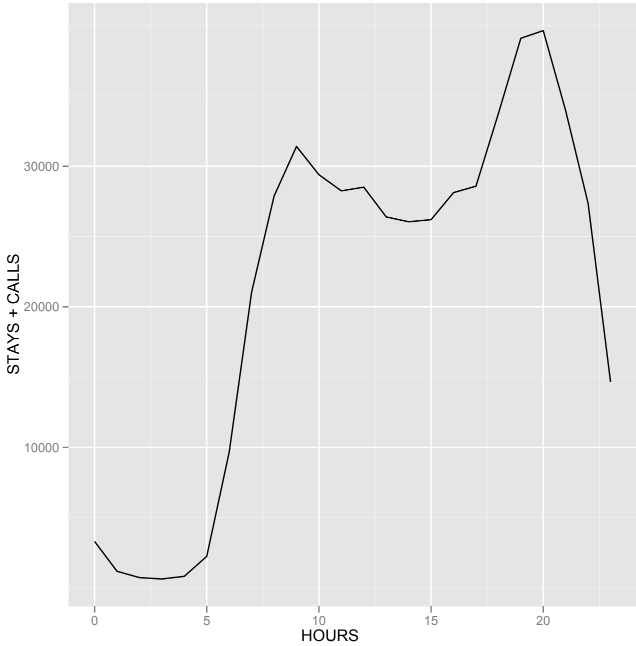
$$\text{S}(p_0, p_1) = \begin{cases} 0 & \text{if } D(p_0, p_1) = 0 \\ 1 & \text{if } D(p_0, p_1) > 0 \end{cases}$$

therefore, the function is defined as follows:

$$N(u, t) = \sum_{i=0}^{n-1} S(\vec{T}_{ut}^{\,i}, \vec{T}_{ut}^{\,i+1})$$

Temporal and geographical components

According to the proposed model, a very important feature has been deduced for the Commuting Dynamic. As we can see in the picture below, there is a couple of time zones when people perform more phone calls from their handsets than usual. Static and dynamic users have not been distinguished, that is, both self-edges and transition edges are counted together.



(a) Total amount of calls grouped by Week days

There are two peaks, p_1, from 7:00 am to 9:00 am, and another one, p_2, from 6:00 pm to 8:00 pm. As we can see, the 'morning-peak' is lower than the 'evening-peak'. Moreover, both peaks reach their maximum value at the end of their corresponding time zone, i.e., p_1 at 9:30 am and p_2 at 7:30 pm. Note how the amount of calls increases linearly since 5 am and how it decreases linearly too since 8:00 pm. Another remarked feature is the existence of a central valley between p_1 and p_2, that is, from 9:00 am to 6:00 pm. This 'peak-valley-peak' pattern is shown for all days in the week, so that we can assume people have the same behaviour (at least, according to our target datasets).

Let's extract some ideas from the previous chart, focusing on what commuters could be probably doing (as we will see shortly, most of the phone calls belong to 'dynamic users'): they get up early in the morning and start to perform more and more phone calls from their handsets until $p1$; next, the amount of phone calls decreases a bit, keeping itself more or less balanced (workhours, lunch) until the beginning of $p2$; later, they people leave the office and plan the rest of the day (errands, leisure...), what is reflected on a marked rise in the amount of phone calls.

Since it is really complicated to measure the displacements of the people (antenna locations instead of users locations, missing antena identifiers...), what will be analyzed is the amount of callers who are really commuters, that is, dynamic users. This one will be the essential magnitude of our research, leading us to figure out the Commuting Dynamics key-features for different regions and times. Eventually, all conclusions deduced from the G/T Model (charts, formulae...) will be ultimately tested with the final GIS visualization as the main core of the work. In this last phase, geographical displacements are estimated so thath thya can be plot in a dynamic interactive map which makes easier to detect peaks, trends...

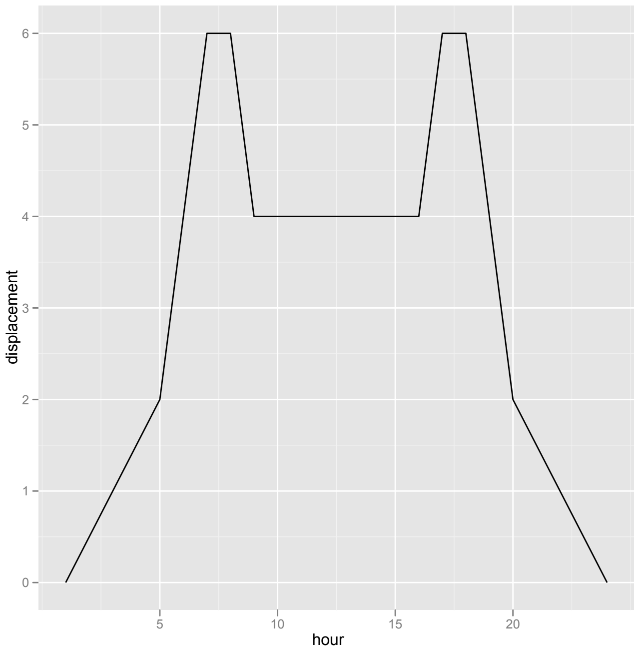
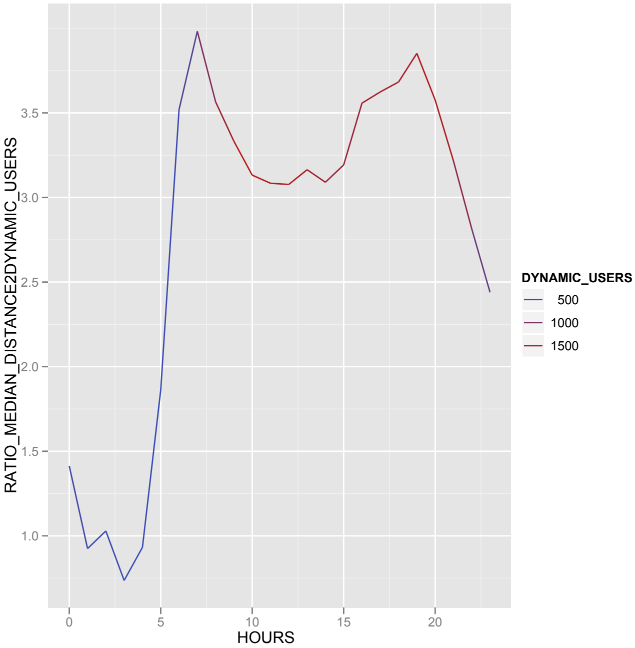


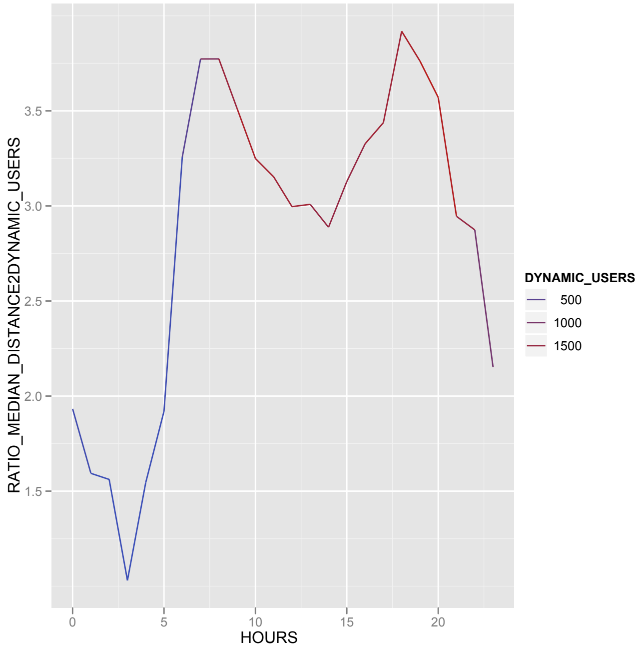
Figure 2: Theoretical Commuting Model

With this new chart, it is the ratio between commuters and all users what it is being emphasized. During a particular day (24h), there is no need to know if a concrete displacement is or not longer than other, no, what it is being highlighted here is the movement detection itself, for 1h windows which groups all users calling within it. The colour of the chart shows how many users are really commuters.

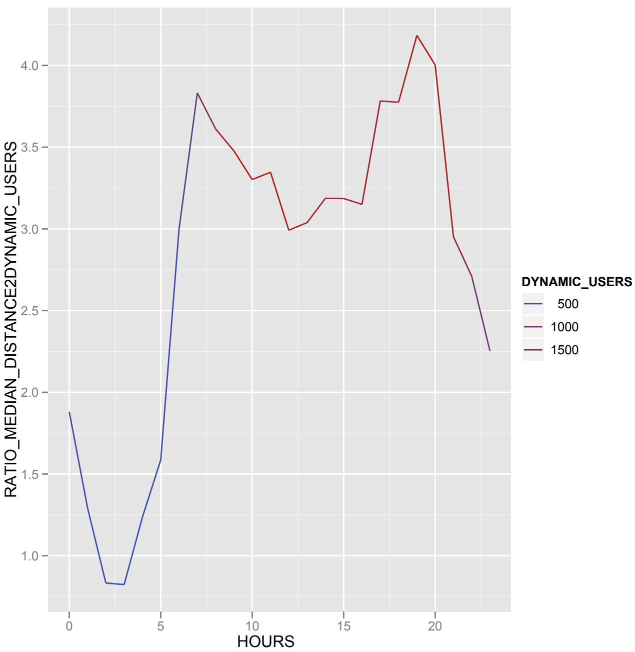
Datasets have been processed according to the proposed G/T Model, filtering non-commuters when commuters tracks have been calculated. Here below, 7 charts display Commuting Dynamics for each week-day. As the results shows, the is not correlation between the number of dynamic users and the maximum displacement peaks, actually, at the same time the number of dynamic users are growing, the central valley an the two maximum commuting peaks are always presented in the sample.



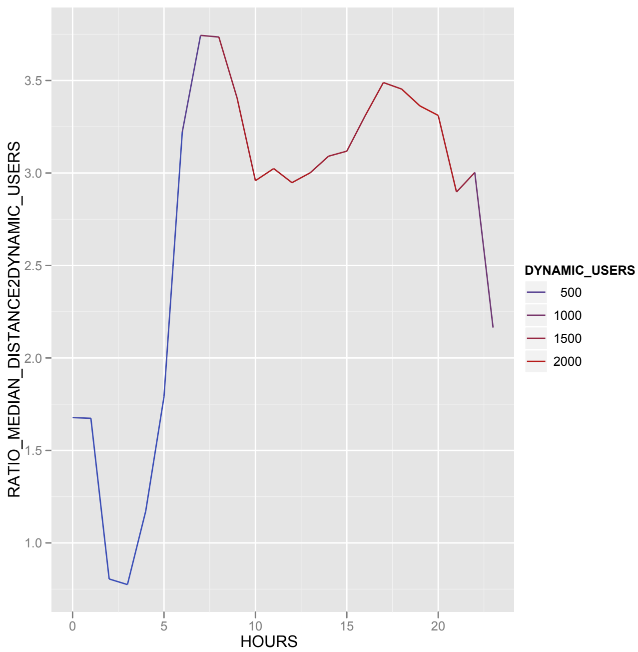
(a) Monday



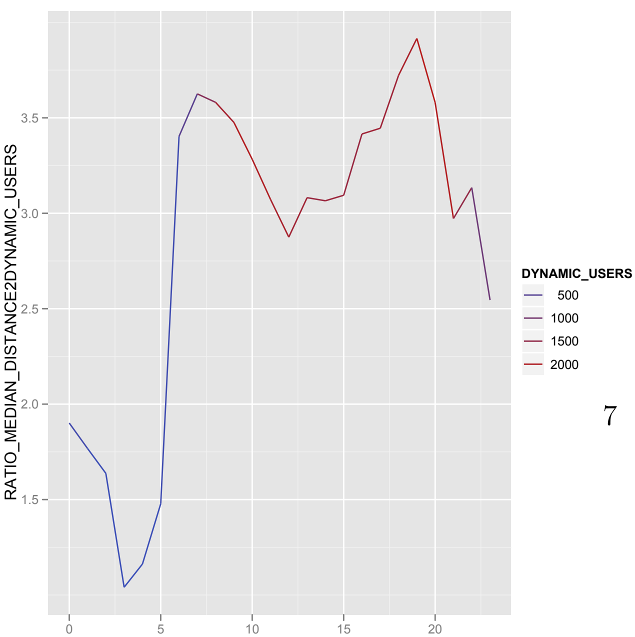
(b) Thuesday

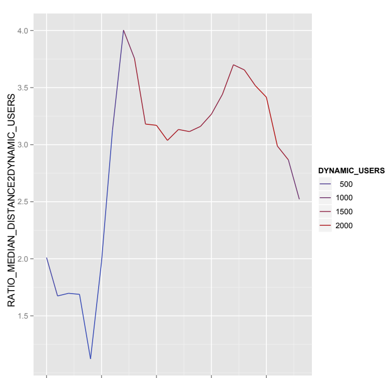


(c) Wednesday



(d) Thursday





Representation and visualization

CSIC, estado del arte de gis, intersección con humanidades, whatever
CSIC/PLRG, Poner los enfoques por los que hemos pasado: los puntos de colores por intensaida de paso, la grilla y finalmente, acabar hablando del kernel density como herramienta final de visualización

To visualize commuting patterns, a diferent approach is used here. To apply a Kernel Density estimation, a geographic network is modeled in order to represent, on the one hand, the antennas positions like nodes on a map and on the other hand, in a time-line order, the user displacements between antennas. With this network, a indegree rank is calculated in order to meassure and set up the weight of each antenna (KDE will use this parameter) and the edge representation is removed from the visualization.

Final Conclusions

Summing up the methods and tools used in this novel study, a mathematical model to shape Space and Temporal user-antenna communications has been proposed, two kind of user dynamics have been identified in order to carry out and focus the study on the commuting dynamics. A commuting pattern model is showed to test with the Orange sample dataset of user communications.

A temporal set of windows defined in 60 minutes ranges and grouped by hours and day of the week has been proposed in the model, however, this assumption could be changed in order to explore the data with another point of view.

Dynamic users has been splitted from statis users to detect commuting patterns, an the mathematical model proposed and appliyed has shown two maximum displacement peaks and a central valley following the main component of total call numbers, normalizing the maximun displacement using the median with the total number if dynamic users.

A visualization of the displacement with a Kernel Density estimation by means of GIS using the antennas positions and users displacements like a direct graph has shown the contraction and the expansion of the commuting dynamics across the 24 hours of the day. A tool to visualizate paralell this dynamic across the seven days of the week is release on line[4].

Poner

Future work

We strongly belive that working with more detailed data i.e.: data from applications and other kind of communications, researchers will be able to work on more accurate mathematical models and therefore, more accurate and useful visualizations. GIS tools are itself a method to research on geolocated data. Several algorithms and methods could be applied to figure out new dynamics and to return the profit of these results into a new develop policys in order to improve the human condition.

[4] "Poner la dirección de las herramientas"

References

A.C. Alegria, H. Sahli, and E. Zimanyi. Application of density analysis for landmine risk mapping. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, pages 223 –228, 29 2011-july 1 2011. doi: 10.1109/ICSDM.2011.5969036.

J. F. Bithell. An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9(6):691–701, 1990. ISSN 1097-0258. doi: 10.1002/sim.4780090616. URL `http://dx.doi.org/10.1002/sim.4780090616`.

John D. Cook. Computing the distance between two locations on earth from coordinates, 2012. URL `http://www.johndcook.com/python_longitude_latitude.html`. [Online; accessed 4-January-2013].

Oscar Marin, Alejandro Gonzalez, Roberto Maestre, Julio Gonzalez, Marco Martinez, Ruben Abad, and Leonardo Menezes. 15th october on twitter global revolution mapped, 2012. URL `http://labs.paradigmatecnologico.com/2011/12/19/15th-october-on-twitter-global-revolution-mapped/`. [Online; accessed 23-January-2013].

Ana Solana and David Alonso. Self-organizing networks and gis tools cases of use for the study of trading cooperation (1400-1800). *Journal of Knowledge Management, Economics and Information Technology*, page pp. 402, 2012. ISSN 2069-5934. URL `http://www.scientificpapers.org/special-issue-june-2012/`.

Wikipedia. Commuting — wikipedia, the free encyclopedia, 2012. URL `http://en.wikipedia.org/w/index.php?title=Commuting&oldid=525136196`. [Online; accessed 4-January-2013].