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

15-Feb-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 findings. 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

List of Figures

About us and 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 final aim has been detecting geospatio-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 to identify new kinds of dynamics.

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 Geospatio-Temporal Model. A Geospatial 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 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 be shown 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 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.*

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

Our first commuting approach is defined like: "Mobility patterns through inferring dynamic users movements grouped by temporal windows".

A *commuter* or *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 a specific hour). Among these temporal windows, *non-commuters* or *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 same user performs these two kind of dynamics within the same temporal window. Note that we are not quantifying the distance, but only the fact of changing from a particular antenna to another one.

State of the art

Nowadays, the world has nearly as many cell phone subscriptions as inhabitants[4]. For the first time, the majority of humanity is linked and has a voice. Consequently, plenty of phone communications are being generated continuously everywhere, and, what is more relevant, they are being tracked: geolocation, start/end times... This is the key, mobile phone companies record data which are very closely associated with behaviour of people.

Analyzing these data in a proper way discloses a great deal of social knowledge (behaviour modeling, people mobility patterns, trends and outliers) which can be applied in countless and different areas[5]: transportation, urban planning, commuting, tourism, traffic congestion, demography, sociology, economy, advertising and commerce, public health... Even without Internet connections (e-mail, IMS and so on), that is, focusing only on speech-calls and text messages, there is a vast amount of information which can be 'read' to reach further conclusions. The ability to understand the patterns of human life by analyzing the digital traces that we leave behind will transform the world, specially poor nations. Reality mining of behavior data is just beginning.

Let's describe a really interesting project [PAPER] about behavioural data. Collecting communication traces into a organization and studying the underlaying patterns, some key outcomes of interest are revealed: social network structure, inferring friendship and proximity levels, individual satisfaction... With temporal data such as call logs, location, phone status, near bluetooth devices, cell antenna ids, application usage(e-mail) and comparing these behavioural data with traditional self-report data show important conclusions.

Regarding D4D datasets (there are only 4 and contain really simple data), note how they have caused many and varied studies from all teams. As far as we are concerned, we discussed about several ideas: antennas network optimization in traffic terms, geospatial-temporal detection of real use for public services (hospitals, schools, police stations...), commuting patterns detection and the like.

Precisely, it has been the human urban mobility approach the one we chose as the core of our project. Why ?? Because it is a reality very tied to ordinary people daily lives, so that its study can reveal clues to improve people quality of life.

Here below a few current researches showing how identified commuting patterns are really useful to understand human motion dynamics better and to perform accurate plans and actions:

a) Exploring spatio-temporal commuting patterns in a Moscow university environment allows making more appropriate decisions to decrease the automobile dependence of students, promoting the non-motorized and public transportation. It is a green initiative looking for sustainability: reducing pollution and noise, avoiding congestion, improving public health and urban planning...

b) Classifying different urban areas based on their mobility patterns from mobile phone data. The results can be used to better understand this dynamic allowing more efficient environmental and transportation policies for the time being and for the future (since due to the regularity of the individual trajectories, it can be claimed that human mobility is highly predictive).

c) Time patterns and geospatial clustering based on mobile phone network data provide accurate statistics about mobility of people, population density and economic activity with detailed regional

[4] http://www.huffingtonpost.com/2012/10/11/cell-phones-world-subscribers-six-billion$_{n1}$957173.*html*
[5] http://www.insead.edu/v1/gitr/wef/main/fullreport/files/Chap1/1.6.pdf

and time resolution.

d) Visual analytics system to study people's mobility patterns from mobile phone data. This tool allows to deeply analyze where, when and who for the calls of people, allowing different kinds of aggregations.

As can be seen, communication data are everywhere (*we are social animals!!*) and they can be used to obtain really interesting and high-value findings. Imagine, once we know the nature and meaning of these data, it is as if we had access to a lot of complete, reliable and immediate surveys. Honestly, we strongly believe that the future lies in knowing how to process this kind of data to get unique results. MIT's Technology Review has recently identified **reality mining on mobile communications** as one of '10 Emerging Technologies That Will Change the World'.

Problem description & Hypothesis

As we saw in the state-of-the-art section, we can extract knowledge from mobile communications datasets. Thus, the solution proposed in this paper is built upon the hypothesis of mobility patterns to predict common and well-known, geographical and time-based models to manage roads and infrastructures in a correlated way with the results figured out.

The figure below shows a theoretical commuting model proposed as a main pattern.
The model shows two peaks, p_1 in the range $[7, 8]$ and p_2 in $[17, 18]$. This first approach to modelize this dynamic sets up the same height for both peaks p_1, p_2. However, numerical results will show that the height of each peak depends on the day of the week. A central valley is defined between $[9, 16]$, with an uniform displacement distribution.
An ad-hoc mathematical model is defined in the next section in order to confirm this hypothesis, consisting on focusing, filtering and processing the main data to contrast the assumption.

The main idea behind the two main peaks, p_1, p_2, and the central valley in the model, is that people cover larger distances in their displacements early in the morning i.e.: p_1 related with the common business activity. After the first peak p_1, people stay in this target destinations, working, eating, etc ..., but in a more static point of view and always performing displacements. The last point in the hypothesis approach shown in Figure 1 is the second peak p_2, when people return to their destination or the last business activites took place.

The geographical behaviour of the proposed dynamic, always mixed with the temporal component, will by contrasted using GIS tools in order to visualize the expansions and contractions in the main points shown by the hyphotesis: p_1, p_2 and the central valley: an expansion when the maximum displacement is reached on the first peak p_1, and a partial contraction when the central valley is reached. Another displacement expansion when peak p_2 is reached and its corresponding contraction around p_2, when it is declining.

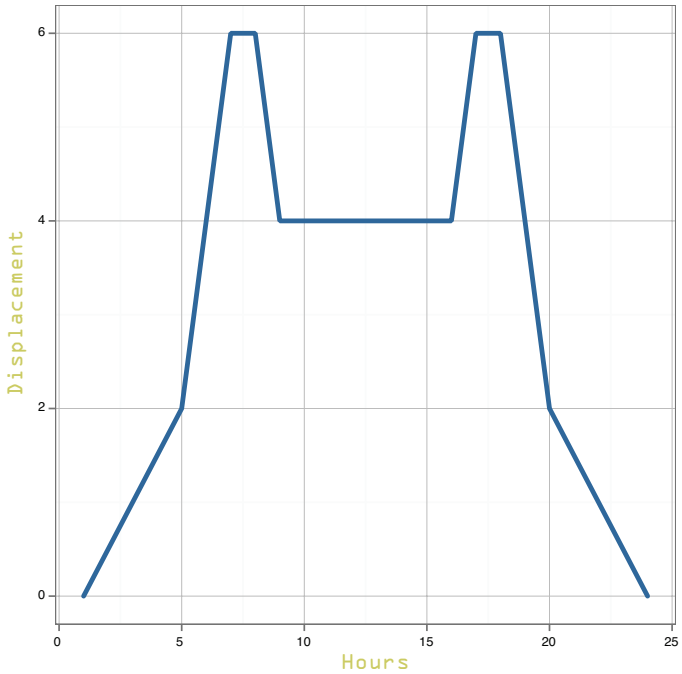


Figure 1: Theoretical Commuting Model

Mathematical model

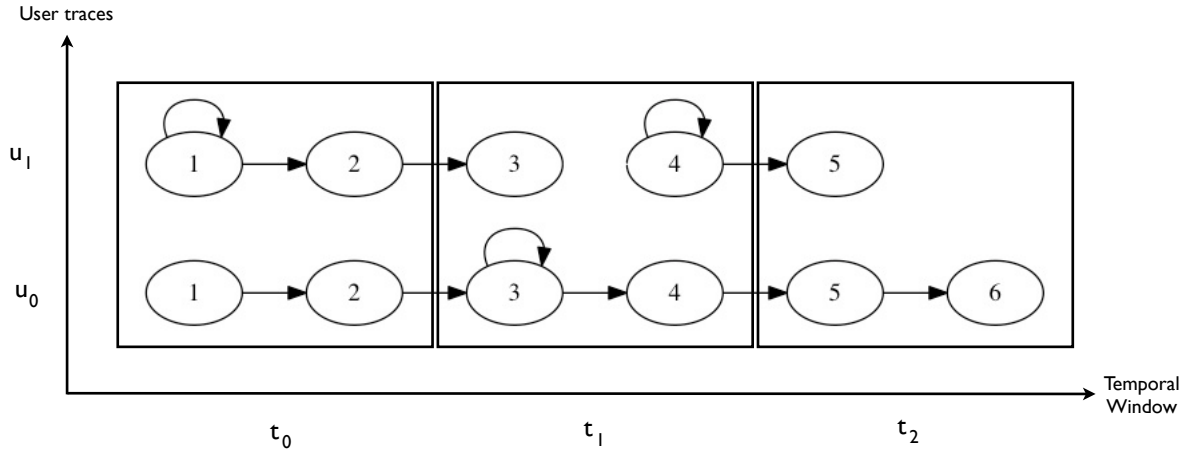


Figure 2: Dynamic and static user patterns

The figure above shows two commuters u_0, u_1 represented on the vertical axis, 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 commuters in a 60 minutes lapse.

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

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

where, $p_n \in \mathbb{R}^2$, $n > 1$, t is a temporal window lapse and, u the unique commuter 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 expressed as 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 follows [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:

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

Methodology

Let's describe how we have faced D4D, enumerating the different phases of our project and highlighting the corresponding milestones. We really think that explaining how this work was carried out can be useful both to better illustrate our conclusions and results, and to give ideas for similar projects.

First of all, once we had clearly understood the D4D bases, we studied all provided datasets to be certain of what kind of data were available. Next, we began with the research work: getting information of Ivory Coast, studying some papers about behavioural patterns obtained from mobile phones traces, looking for new related datasets...

With this knowledge, we were prepared to decide which lines of work would be more interesting (without forgetting the cooperative and development goal apart from the scientific one) and, what's more important, being aware of our own time constraints and our team skills –being realistic is crucial.

After some discussion, we agreed to focus on the 2nd dataset 'Individual Trajectories: High Spatial Resolution Data (SET2)' [LINK], since it seemed to be the most adequate one for our approach. We conducted our analysis according to the following stages:

1) Processing all traces, grouping them by user and sorting them chronologically, as hourly time series. Paying attention to imprecise or weird traces, which must be filtered.

2) Calculating different magnitudes (absolute, relative and normalized ones) and their mean, median and dispersion to display visual charts, which helped us to discover correlations and to identify 'Temporal Commuting Patterns' for each week day. [GRAFICA]

3) Handling antenna locations from the previous processed traces, allowed us to identify 'Geospatial Commuting Patterns'. Firstly, we represent networks graphs and some static maps (snapshots of commuters motion). Later, we were able to create animated and detailed maps (Kernel Density, Grids...) which made easier to see crowded areas, related highways... during the days and all across Ivory Coast. [GRAFICA]

4) Eventually, an online and interactive web-based animation was developed. This geovisualization technique is advantageous in that neither specialized GIS knowledge nor software is required, and it enables change over time visualization that would be difficult to see with static or paper maps. The interface combines raster maps produced in the ArcGIS environment and vector data [PANTALLAZO]. User interaction is facilitated through the inclusion of buttons on the interface (play controls, modal tab, zooming and panning).

As can be seen, the whole process to obtain the results has been carried out step by step. We had a planning which was useful, but the really important thing was the fact of planning, not the planning itself. There will be unexpected events which required the team to adapt itself to new circumstances.

In the end, we would like to remark how, although assigning particular tasks to different team members looking for productivity, all of us have tried to be involved in all areas.

Results

Here below a brief enumeration of the achieved results during this project.

a) A designed and implemented mathematical model to detect geospatial-temporal commuting patterns.

b) A set of charts and maps which illustrate the previous model, making easier to deduce interesting findings.

c) An on-line application[6] to display all this information in a friendly and customizable way.

Now, let's describe deeply them, as well as other partial findings.

According to the proposed model, a very important feature has been deduced for the Commuting Dynamic. As seen in the picture below, there are two 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.

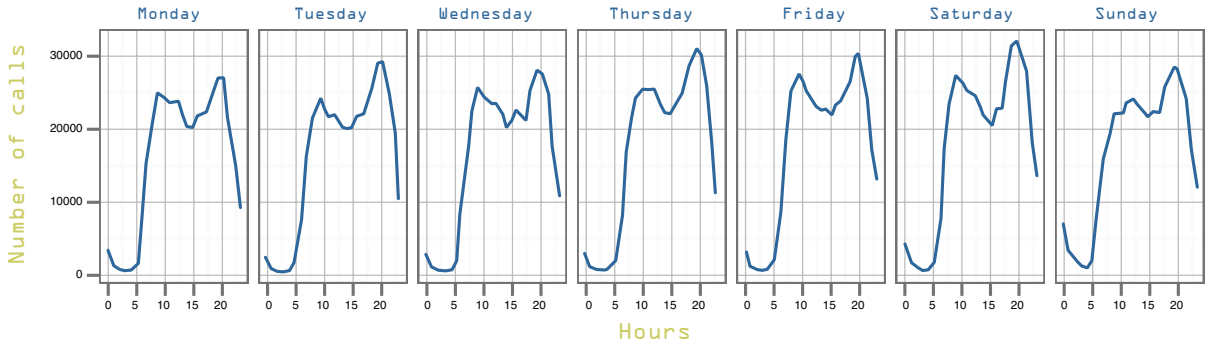


Figure 3: Total amount of calls grouped by Week days

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 it can be seen, the 'morning-peak' is lower than the 'evening-peak'. Moreover, both peaks reach their maximum value at the end of their corresponding time zones, 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 remarkable 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 it can be assumed that 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 it will be seen 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 slightly, keeping itself more or less balanced (workhours, lunch) until the beginning of $p2$; later, 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

[6] "Poner la dirección de las herramientas"

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 that they can be plotted in a dynamic and interactive map which makes easier to detect peaks and trends.

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. What it is being highlighted here is the movement detection itself, in 1h windows which group all users calling within them. 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, seven charts display Commuting Dynamics for each week-day. As shown in the results, there is no 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 and the two maximum commuting peaks are always presented in the sample.

These seven charts below show a fitter correlation with the theoretical commuting model proposed in this paper (Figure 1), displaying two high peaks and a lower central valley. However, more qualitative data is necessary to figure out the performance of the first high peak, because it is not related to the number of dynamic users. Furthermore, a first approach can be applied to say that few dynamic users (in comparision with the mean) travel longer distances in this first peak, specially in the first uphill to the first peak. This is a common pattern figured out from the proposed model. However, it can be explained because there are people starting their travel from further distances early in the morning, since the most of the people live near their work places and travel nearer distances than the first group.

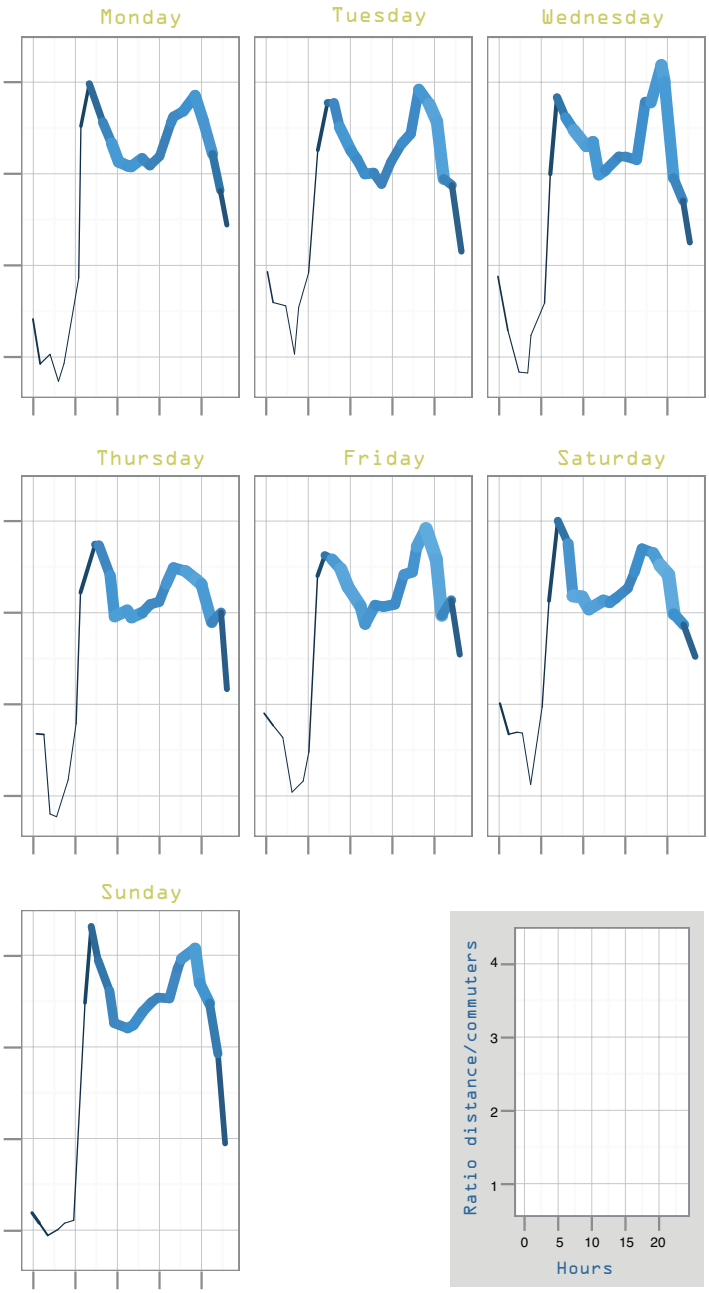
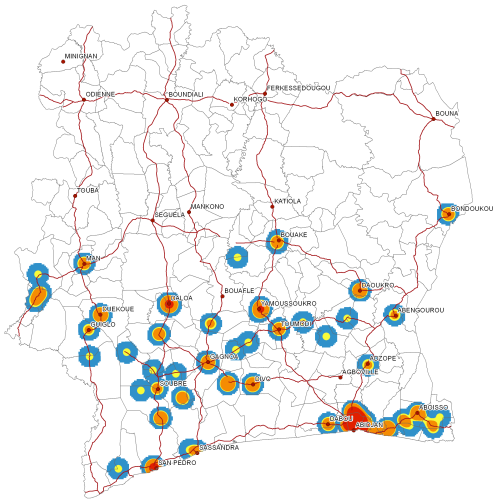
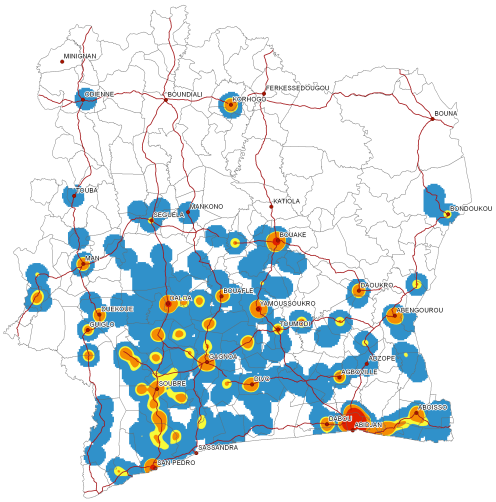


Figure 4: Dynamic users displacements

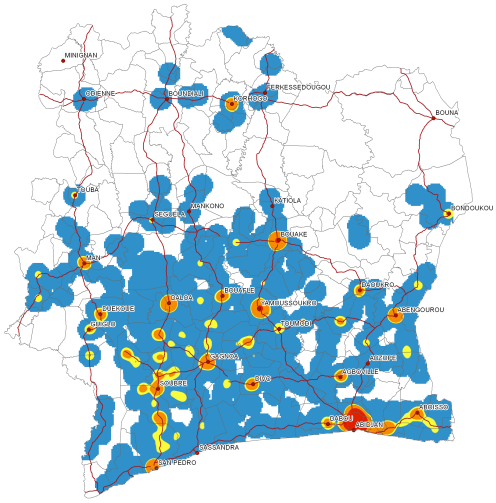
A visualization of the displacement with a Kernel Density estimation using GIS with the antennas positions and users displacements like a directed graph has shown the contraction and the expansion of the commuting dynamics across the 24 hours of the day. The firsts expansion visualization through the kernel density estimation is clearly showed when peak p_1 starts to grown between $[5, 9]$.



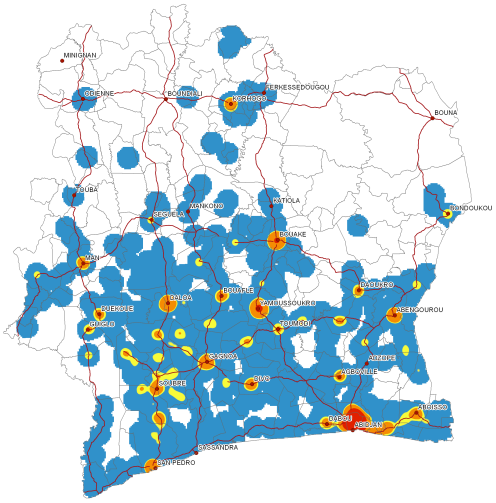
(a) Monday, 5 AM



(b) Monday, 6 AM



(c) Monday, 7 AM



(d) Monday, 8 AM

Figure 5: Evolution of KDE while peak p_1 is reached

Conclusions

After having completed the project, we have demonstrated with a practical example how relevant conclusions can be extracted by analyzing this kind of mobile phone networks data. In particular, a couple of general ideas have been confirmed:

a) Mobile phone traces are all over the world and they hold plenty of high value information.

b) Among possible behavioural patterns extracted from a), knowing people dynamics, and specially commuting patterns, constitutes a valuable tool to improve infrastructure and public services both for the time being (detecting crowded/empty areas or periods) and for the future (predictions).

Focusing on D4D, we could remark these partial findings:

a) Distinction of *commutters* and *noncommuters* groups and their evolution during each week day and for different Ivory Coast regions.

b) Guessing of common usage of mobile phones, in amount of calls terms, during a day and in different cities.

c) Identification of a morning peak (9:30am) and an evening peak (7:30pm), in maximum displacement terms, values have been normalized using the median of total amount of commuters.

d) Identification of lower valley between both peaks.

As the main conclusion for this project, it could be claimed that, basing on the collected Orange mobile phone traces in Ivory Coast during the observed period of time, **a couple of commuting peaks have been identified for each day of the week, with a more defined pattern for work days and for big cities (Abidjan, Bouaké, Daloa, Yamoussoucro)**. Moreover, **people motion between the outskirts and the city center in the morning, and vice versa in the evening could be detailed for each particular city, highlighting those road segments with more traffic**.

Apart from the 'commuting' conclusions, there is another one which deserves to be exposed. After some discussions, we decided to create an accesible, user-friendly and customizable tool so that this kind of data could be actually profitable. How can you expect this complex process can be understood for some common people if you do not make things easy ?

Summing up, we hope this global D4D effort can help Ivory Coast in decision making for policy measures and ultimately leads to perform some trustworthy plans to improve *ivoriens* quality of life. Taking advantage of a tool like this one will save them money and time.

Recommendations for further work

This project has reached interesting results, but it can still be evolved in several ways. Due to not having enough time, some ideas were only proposed and they could not be developed. They have been listed here below as an outline for those teams who agreed with us in seeing their potential related to the model.

a) **Particularization**: general findings are useful to know how to face a problem initially, but with concrete findings more adequated solutions will be reached. Specific conclusions and maps for different cities and regions could be calculated.

b) Clustering and/or filtering antennas: by traffic (using the 1st dataset), by location (city/field, latitude)...

c) Replicating the model and results with the 3rd dataset.

d) A 60-minutes time span and a daily displaying approach have been used. However, both assumptions could be modified in order to explore data from another perspective: season patterns, overlapped or shorter time spans...

e) Looking for more understandable charts: normalizing time series by dividing all values by the maximum of the day... and some other strategies to obtain relative magnitudes.

f) Trying to find correlation evidence between the amount of calls rate and prosperity indicators (business, grants...).

g) Looking for additional socio-economic datasets to conduct new mixed analyses (weather...).

h) Developing a new kind of maps based on tessellations (i.e., Voronoid diagrams), which have been proved to be really useful for this kind of studies.

i) Applying DTW[7] or LCS[8] algorithms to discover similarities among different traces series, in order to identify types of areas (residential, commercial, business) and to predict its evolution in mobility terms.

j) Trying to identify where people live and work.

k) **Outliers detection**: both in particular time zones and in specific regions, or also regarding patterns evolution and consolidation.

Moreover, some ideas and modules of this project are hoped to be used into another completely different works, so that original and novel conclusions are found with a high synergy value.

Eventually, we strongly believe that working with more detailed data (both Internet and Call/SMS communications) will let researchers find more accurate mathematical models and therefore, more precise and useful visualizations. Furthermore, having longer time span datasets available would assert the patterns and would allow to describe more reliable people dynamics trends.

GIS tools are themselves a fantastic way to research on geolocated data. Applying different GIS

[7] http://en.wikipedia.org/wiki/Dynamic$_{t}ime_{w}arping$

[8] http://en.wikipedia.org/wiki/Longest$_{c}ommon_{s}ubsequence_{p}roblem$

algorithms and methods could obtain new dynamics patterns, providing profitable new results to be taken into account when better development policies are decided in order to improve the people quality of life.

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].