D4D Challenge
Commuting Dynamics 4 Change

R. Maestre[1], R. Lario[1], M. Muñoz[1], R. Abad[1], J. Gonzalez[1], A. Martín[1], E. Perez[2], and JL. Fdez-Pacheco[3]

[1]Paradigma Labs (PL) , Paradigma Tecnológico (PT) ,
rmaestre@paradigmatecnologico.com
[2]GIS Laboratory , Centre for Humanities and Social Sciences (CCHS) , Spanish National Research Council (CSIC) , esther.perez@cchs.csic.es
[3]Complutense University (UCM) , jlfpsaez@trs.ucm.es

15-Feb-2013

Abstract

Our idea is to use the geolocation data from the antennas processing the mobile phone calls in order to know which sub-prefectures the customers have been getting around. The main goal of our project is developing geospatio-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, etc. 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 Ivory Coast surface, working with geographical layers: landcover, roads map, railway lines, water sources, etc. Consequently, the ched 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, etc.

Keywords: Human Dynamics, Commuting, GIS, Data Analysis, Mobile Communications.

Contents

List of Figures

Commuted Communities: Commuting Dynamics 4 a Change

For decades, Africa has been receiving official development assistance, ODA, from the so-called countries of the "North". There have been several models of cooperation with this continent, from bi-lateral country agreements passing through religious and non-governmental organizations to new models of financing development. None of them is free from dark sides and controversial issues in this "culture of aid" (Moyo, 2009). The African continent doesn't need that kind of helping agreements which submerge the continent in a dark hole and position it on unequal relations with international financial institutions which pretending good wishes create difficult conditions that keep on enslaving the continent on new ways of Neo-colonialism. African countries are in need of endogenous conditions for creating their path to democracy and their own government empowering. We understand development as a process linked to the capacity of African people to decide about its own future(SEN et al., 2000). In this sense, our proposed ways of cooperation pass through the vision of an horizontal structure where sharing-knowledge conditions are created in order to allow the empowering in every direction. As the one we introducing on this article, our proposals are based on the new technologies and the way they can help to develop new aspects as far as planning and researching are concerned in any country they are implemented.

The emergence of new technologies and virtual tools in the field of development cooperation had an early stage in the late nineties(De Jong et al., 2005). In that period, the developed projects were aimed at providing hardware structures and networks to disadvantaged communities or with communication difficulties due to the orography within their territory. Nevertheless, our proposals on this article are based on software tools that have been already experienced in countries such as Ivory Coast where the development conditions are allowing to start testing some of these tools.

GIS based technologies have been sufficiently proved in some sustainable development projects carried out in Africa[1] related to international cooperation as well as in local development where the cross-sectorial work is critical and the shortage of funding make it, sometimes, impossible (Mitchell, 1997), (Craig et al., 2002). Recently the GIS has a crucial role in the democratic processes in Africa, as we can see in the creation of the Census 2012 in Rwanda where the GPS data collection will enable them to build a comprehensive GIS database, which shows boundaries, location of schools, hospitals and markets covering 17,700 villages[2].

Also in South Africa, part of our team had the experience of working with GIS as a tool to identify agricultural and rural areas where to start community vegetable gardens as well as managing the cattle in some projects implemented in KwaZulu-Natal region.

As we said, the two tech experiences gathered on this article look deeply their contribution to serve as tools to facilitate the cooperation in two aspects:

a) It will allow to develop better ways to plan and manage the territory since their GIS based structure will permit us to work with geographical layers: landcover, roads map, railways lines, water sources, etc. Consequently, this information will allow to visualize and reach some conclusions in order to design and formulate better cooperation and local development projects based on the municipalities.

b) This tool will become an useful resource in social research(Fielding et al., 2008) on deterritorialized and transnational realities(Sassen, 2007) since it brings new possibilities of comparative analysis and the creation of resources ad hoc.

---

[1]http://www.esri.com/library/brochures/pdfs/gis-for-africa.pdf
[2]http://www.humanipo.com/news/1279/GIS-based-enumeration-kicks-off-in-Rwandas-2012-census

About us and why we face this challenge

Paradigma Labs[3] (PL) 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 and improving quality of life for people, quickly held our attention.

GIS laboratory[4] at Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council[5] (CSIC) is a multidisciplinary group with a huge experience in Remote Sensing and Geoprocessing, providing a quality support for plenty of research projects carried out at CSIC.

Our final aim has been detecting geospatio-temporal patterns in order to obtain 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 other 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 comunications', 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 such as 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 dynamics, not only moving across the timeline, but 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 community, allowing to study the dynamics more deeply. 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 prove which and where the maximum commuting peaks are.

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.* Our first commuting approach is defined like: "Mobility patterns through inferring dynamic users

---

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

movements grouped by temporal windows".

A *commuter* or *dynamic user* is defined as an user changing their antenna location within the studied temporal window (i.e.: each temporal window groups the whole user communication for 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 the same user performs these two kind of dynamics within the same temporal window.

State of the art

Nowadays, the world has nearly as many cell phone subscriptions as inhabitants[6]. 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[7]: 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 behaviour data is just the beginning** .

Let's describe a really interesting project (Eagle et al., 2009) about behavioural data. By collecting communication traces into an organization and studying the underlying patterns, some key outcomes of interest are revealed: social network structure, inference of friendship and proximity levels, individual satisfaction... This is achieved with temporal data such as call logs, location, phone status, near bluetooth devices, cell antenna ids, application usage(e-mail). Comparing these behavioural data with traditional self-report data show important conclusions.

Regarding D4D datasets (there are only 4 and they contain really simple data), note how they have caused many and varied studies from all teams. As far as we are concerned, we discussed 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 alike.

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

Below, we can see a few current research projects 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 these dynamics 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 predictable**).

c) Time patterns and geospatial clustering based on mobile phone network data provide accurate

---

[6]http://www.huffingtonpost.com/2012/10/11/cell-phones-world-subscribers-six-billion_n_1957173.html
[7]http://www.insead.edu/v1/gitr/wef/main/fullreport/files/Chap1/1.6.pdf

statistics about mobility of people, population density and economic activity with detailed regional 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 from the calls of people, allowing different kinds of aggregations.

As it 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'.

Several studies have proven the utility of Geographic Information Systems (GIS) for commuting analysis (flows intensities and directions), because of its efficiency dealing with data with a geographic component. This is the case of people displacement patterns on a particular portion of space. That, combined with its ability to represent data over the territory through geovisualization techniques, makes GIS one of the most used tools for this kind of studies.
The conclusion of all of those studies points to the valuable information that can be extracted from commuting, in terms of city growth, decisions about cities and companies placement or people disposition to migrate or dealing with long journeys to their job or leisure places.

There are many applications and use cases in the bibliography, including: the analysis of work commuting in a city, in terms of time and distance to locate spatial variations (Wang, 2000), the analysis of people behaviour patterns related to their leisure activities, the study of georeferenced commuting patterns to elaborate models that predict workplace contacts which result in disease transmission (Chrest and Wheaton, 2009), the study of dispersión trends or concentrations in specific studied areas, the study of average distances spent on displacements as well as spent times relating them with companies and city location patterns, the study of the differences between mobility patterns of national and immigrant employees (Llano Verduras, 2006), the evaluation of optimal routes (Thériault et al., 1999) the analysis of costs and transport problems in intraurban and interuban structures (Zhan et al., 2008), etc.

Problem description & Hypothesis

As we saw in the previous section, we can extract knowledge from mobile communication datasets. Thus, the solution proposed in this paper is based 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 (theoretical commuting model proposed as a main pattern) shows two peaks, p_1 in the range 07:00-08:00 and p_2 in 17:00-18:00. This first approach to modelize these dynamics 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 09:00-16:00, 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 these 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 dynamics, always mixed with the temporal component, will be 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.
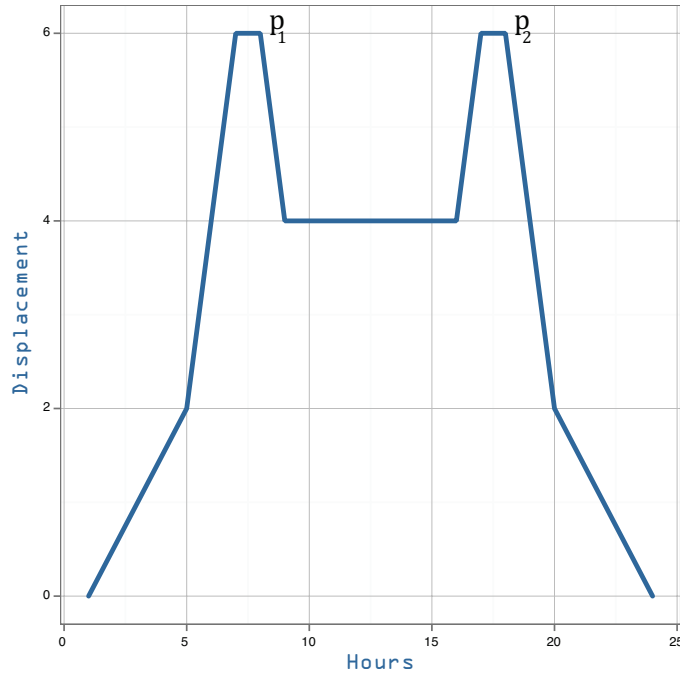

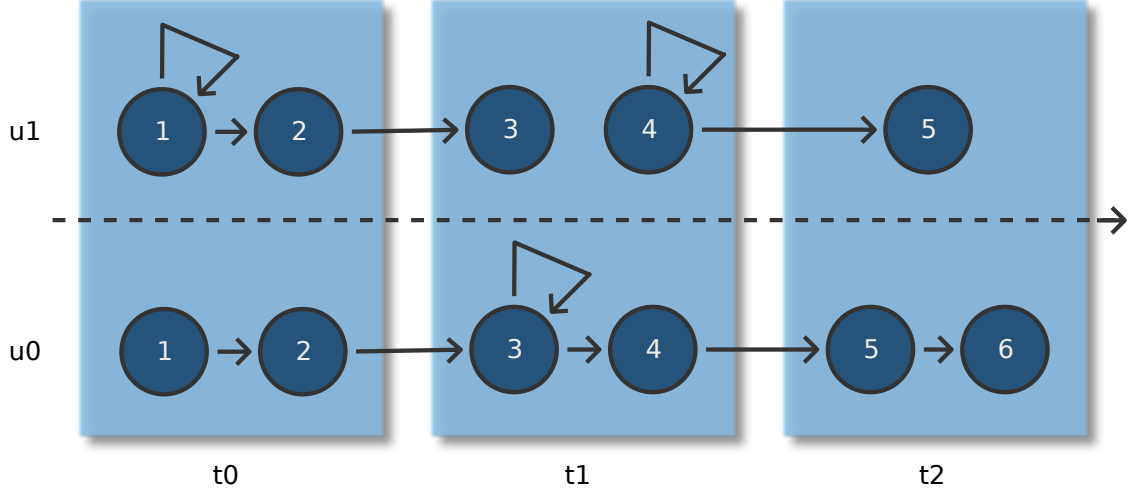Figure 1: Theoretical Commuting Model

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 communication traces of the whole set of commuters in a 60-minute lapse.

Let $\vec{P_i} = (p_{ix}, p_{iy})$ be the position vector where p_{ix} is related tos the geographic latitude and p_{iy} is related to the geographic longitude of each position.

Formally, a commuter trace during a particular time window and related with a specific user is defined as follows:

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

where, $\vec{p_n} \in \mathbb{R}^2$ and $n > 1$.

For instance, regarding with user 0 and time window 0 we have the next user trace $T = (\vec{p_1}, \vec{p_2})$. Regarding with user 1 and time window 0 we have the next user trace $T = (\vec{p_1}, \vec{p_1}, \vec{p_2})$ 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:

$$D(\vec{p_0}, \vec{p_1}) = \mathrm{acos}(\sin(\phi(p_{0x})) * \sin(\phi(p_{1x})) * \cos(\theta(p_{0y}) - \theta(p_{1y})) + \cos(\phi(p_{0x})) * \cos(\phi(p_{1x})))$$

where:

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

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

therefore the function related to the distance and regarding to a specific user into a temporal window is defined as follows [result in Km]:

$$U = 6373 * \sum_{i=0}^{n-1} D(\vec{p}_i, \vec{p}_{i+1})$$

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:

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

therefore, the function regarding a specific user into a temporal window is defined as follows:

$$N = \sum_{i=0}^{n-1} S(\vec{p}_i, \vec{p}_{i+1})$$

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)' (Blondel et al., 2012) , 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.

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.

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. 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[8] to display all this information in a friendly and customizable way.

Now, let's describe them deeply, 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.
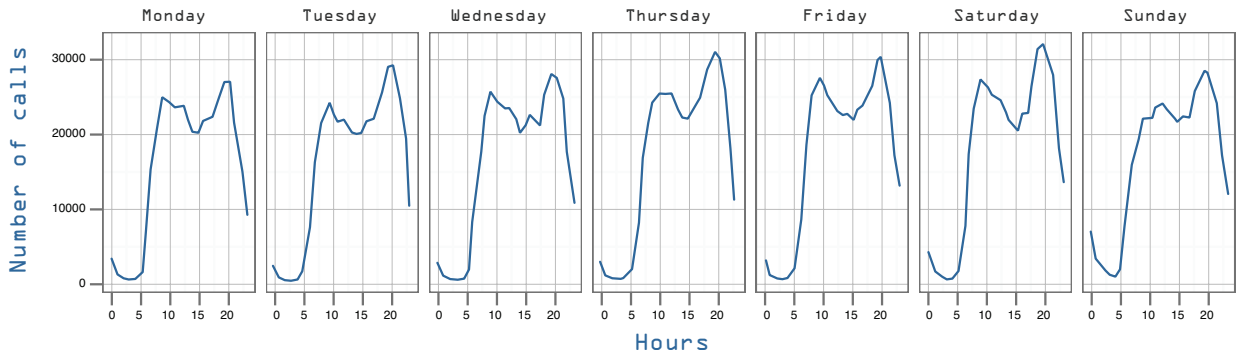

Figure 3: Total amount of calls grouped by Week days

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 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.

---

[8] http://labs.paradigmatecnologico.com/d4d/

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 are 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 a 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.
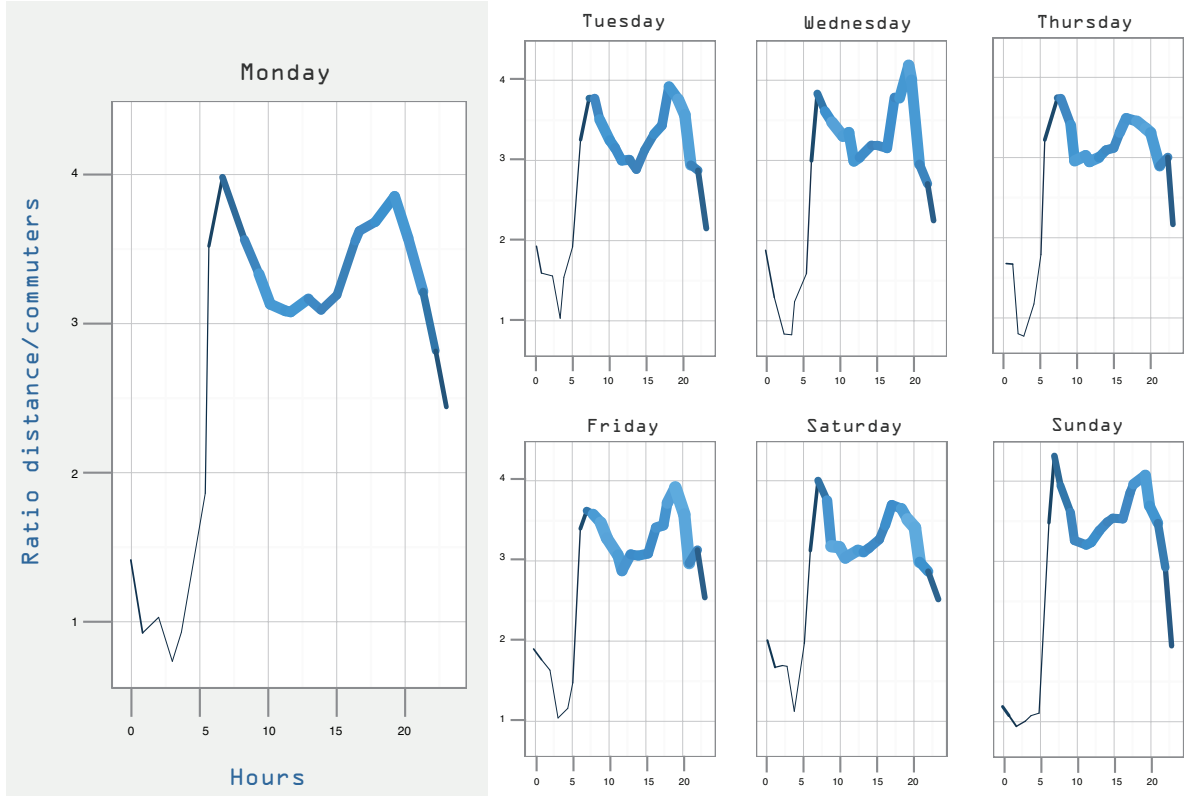

Figure 4: Dynamic users displacements

Density analysis takes known amounts of some phenomena and spreads them across the landscape basing on the quantity that is being measured at each location. With the position of the antennas of mobile telephony and the estimated weight for each of them, depending on their traffic intensity, a kernel function has been used to calculate the density of features in a neighbourhood around those features. It calculates a magnitude per unit area from point features (antennas) using a kernel function to fit a smoothly tapered surface to each point. The result is a surface, a raster image. The original information of the antenna positions was expressed in geographical spherical coordinates (longitude, latitude), but the kernel functions needs to work with projected information to calculate the distances, therefore a previous conversion of the same ones was performed projecting them to a projection system UTM using the datum WGS84 and the zone 30N. Kernel Density calculates the density of point features around each output raster cell; the kernel function is based on the quadratic kernel function described in Silverman, 1986. Conceptually, a smoothly curved surface is fitted over each point. The surface value is higher at the location of the point and diminishes with increasing distance from the point, reaching zero at the 'search radius' distance from the point. Only a circular neighborhood is possible. The volume under the surface equals the 'population field' value for the point. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center. The cells nearer the measured points, the antennas weight, receive higher proportions of the measured quantity than those farther away. (ESRI©, 2012).
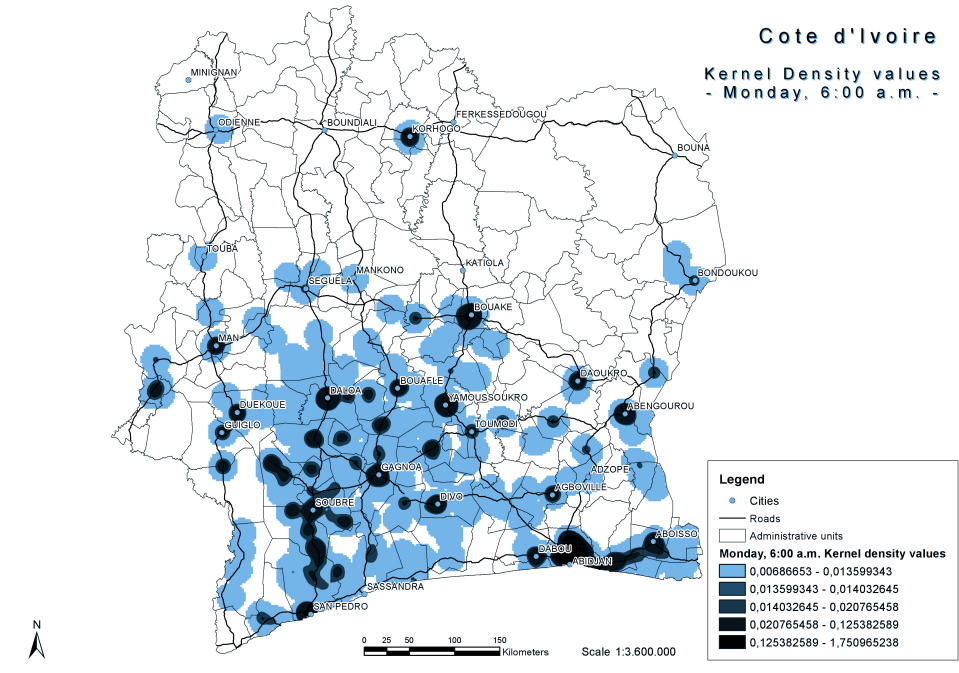

Figure 5: Kernel Density sample

A visualization of the displacements 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 grow between 05:00 and 07:00 reaching the central valley at 8:00.
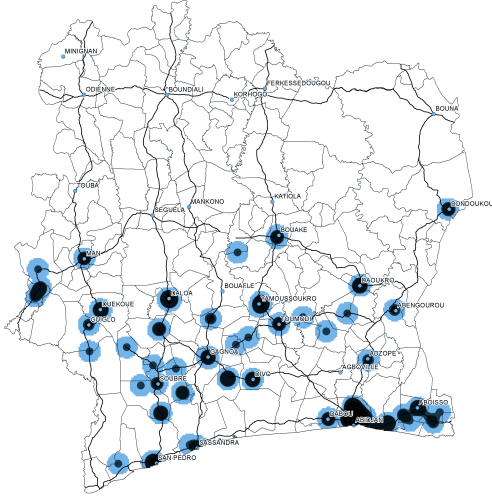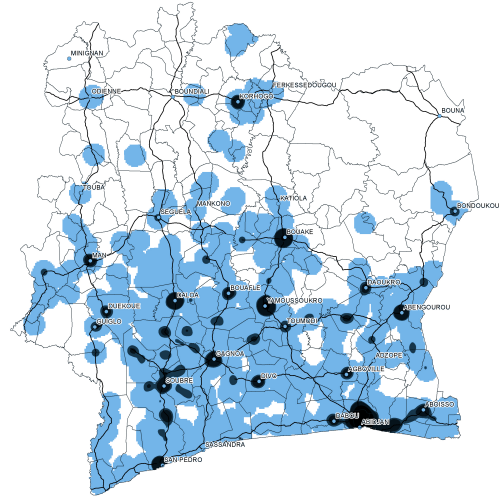
(a) Monday, 05:00
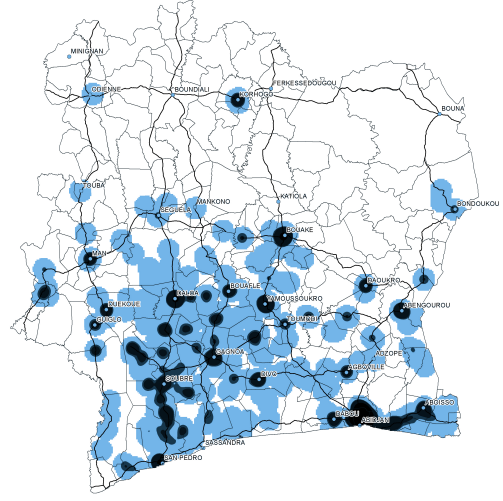
(b) Monday, 06:00
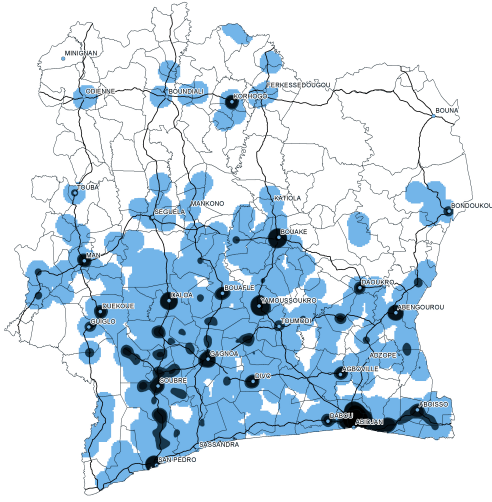
(c) Monday, 07:00

(d) Monday, 08:00

Figure 6: KDE evolution while peak p_1 is reached

Also, while a normalized displacement meassure is distributed across the country in the valley hours, a reduction of the displacement of the dynamics users can be appreciated especially on roads where KDE had a high intensity. Above picture, Monday 06:00, shows a high intensity kernel distributions, coinciding with the maximum displacement moment calculated and showed in the Figure 4 on Monday. These values should be identified regardless each week day commuting displacement distribution, however, the ranges of peaks p_1, p_2 and the central valley are leading for a regular temporal pattern.

Next pictures shows the reduction on the KDE for the roads among main cities in a more detailed way. We have selected a sample zone in order to show the correlation between the numeric distance calculated by means of the proposed mathematical model and the KDE calculated by means of GIS technology, to test the final aim proposed in this study.
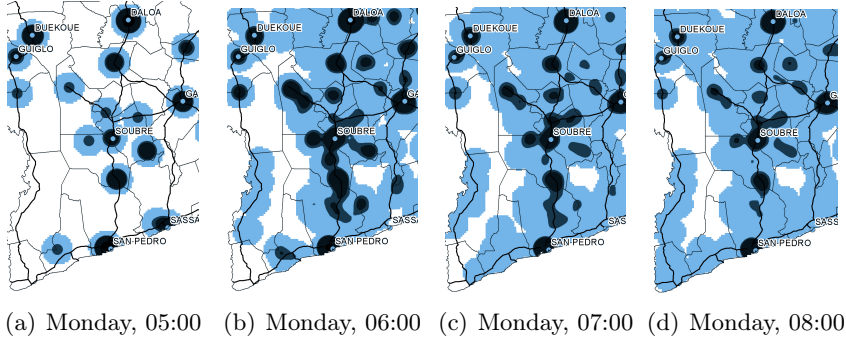
16

(a) Monday, 05:00 (b) Monday, 06:00 (c) Monday, 07:00 (d) Monday, 08:00

Figure 7: KDE focus on p_1

Therefore, the pictures above show the evolution of KDE in the target location when maximum displacement p_1 is reached at 06:00, after this hour, KDE becomes softer at 07:00 and 08:00 starting the central valley. Through KDE we can figure out that the median displacement in the central valley 08:00 and go on until p_2 has a greater value as numerical results leads.

The central valley has a uniform distribution of KDE and also in the next picture we can observe how the intensity increases when peak p_2 is reached at 19:00. We propose the same location to explore and correlate the numerical results of p_2 with the final visualization of the dynamic taking into account the two variables.
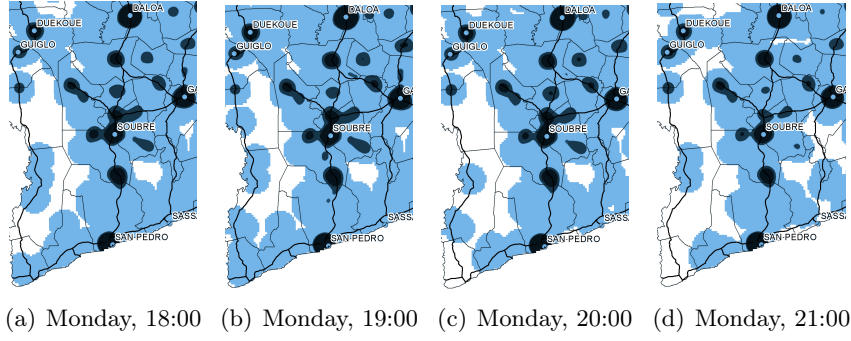

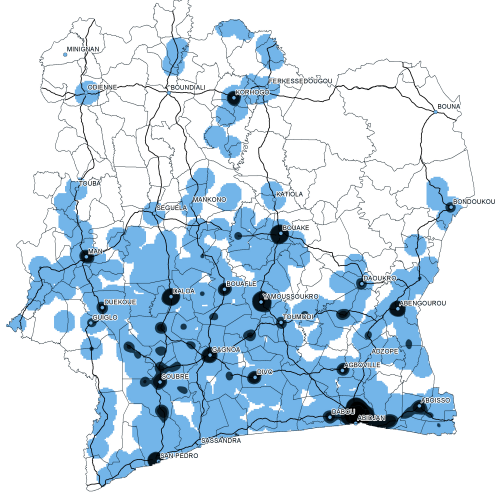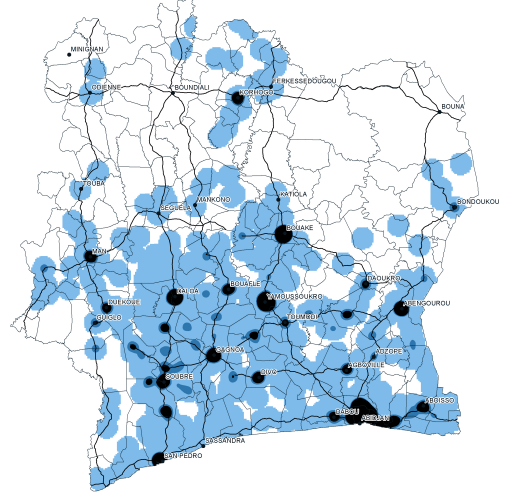(a) Monday, 18:00 (b) Monday, 19:00 (c) Monday, 20:00 (d) Monday, 21:00
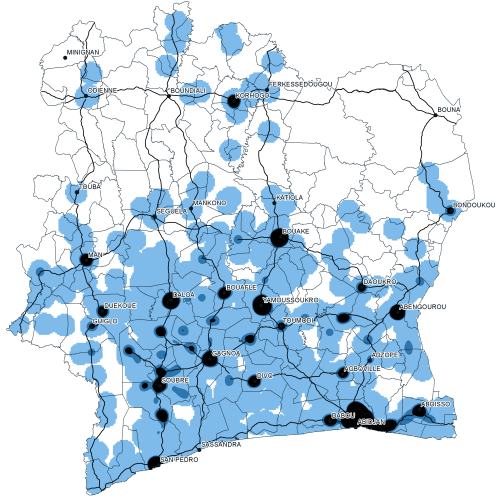
Figure 8: KDE focus on p_2

In a more detailed way, the whole extension picture is provided by the next four pictures representing the KDE of dynamic users across the country.
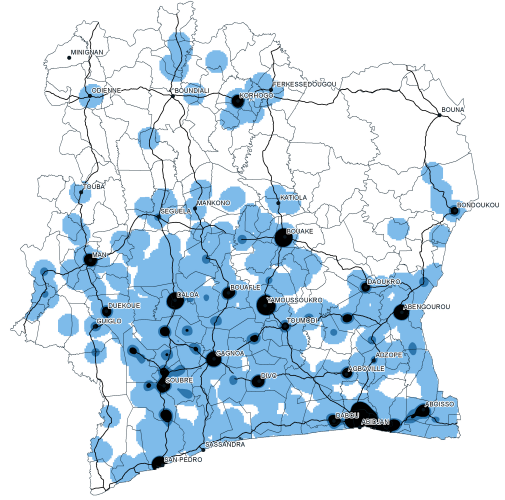
17

(a) Monday, 17:00

(b) Monday, 18:00

(c) Monday, 19:00

(d) Monday, 20:00

Figure 9: KDE evolution while peak p_2 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 behavioral 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 and plannings).

Focusing on D4D, we could remark these partial findings:

a) Distinction of *commutter* and *noncommuter* 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 (09:30) and an evening peak (19:30), in maximum displacement terms, values have been normalized using the median of total amount of commuters.

d) Identification of a lower valley between both peaks.

As the main conclusion for this project, it could be claimed that, based 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 to be understood by some 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 will lead to perform some trustworthy plans to improve the quality of life of the *ivoiriens*. Taking advantage of such tool will save them money and time.

Recommendations for further work

This project has reached interesting results, but it can still be developed 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 introduced model.

a) **Particularization**: general findings are useful to know how to face a problem initially, but with concrete findings more adequate solutions will be found. 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), etc.

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

d) A 60-minute 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 values by the maximum volume 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, etc.).

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

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

i) **Applying DTW[9] or LCS[10] algorithms** to discover similarities among different traces series, in order to identify types of areas (residential, commercial, business) and to predict their 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 expected to be used into another completely different fields, so that original and novel conclusions can be reached with a high synergy value.

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

---

[9]http://en.wikipedia.org/wiki/Dynamic_time_warping
[10]http://en.wikipedia.org/wiki/Longest_common_subsequence_problem

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

For instance, a traffic density estimation through user dynamics comunications is proposed. The next picture shows an aggregated value, however, a flow intensity related to the direction of each user trace can be applied too. A correlation of segment creation from each two antenna positions and the flow intensity between antennas taking into account the direction, could be directly applied to a real time traffic manager.
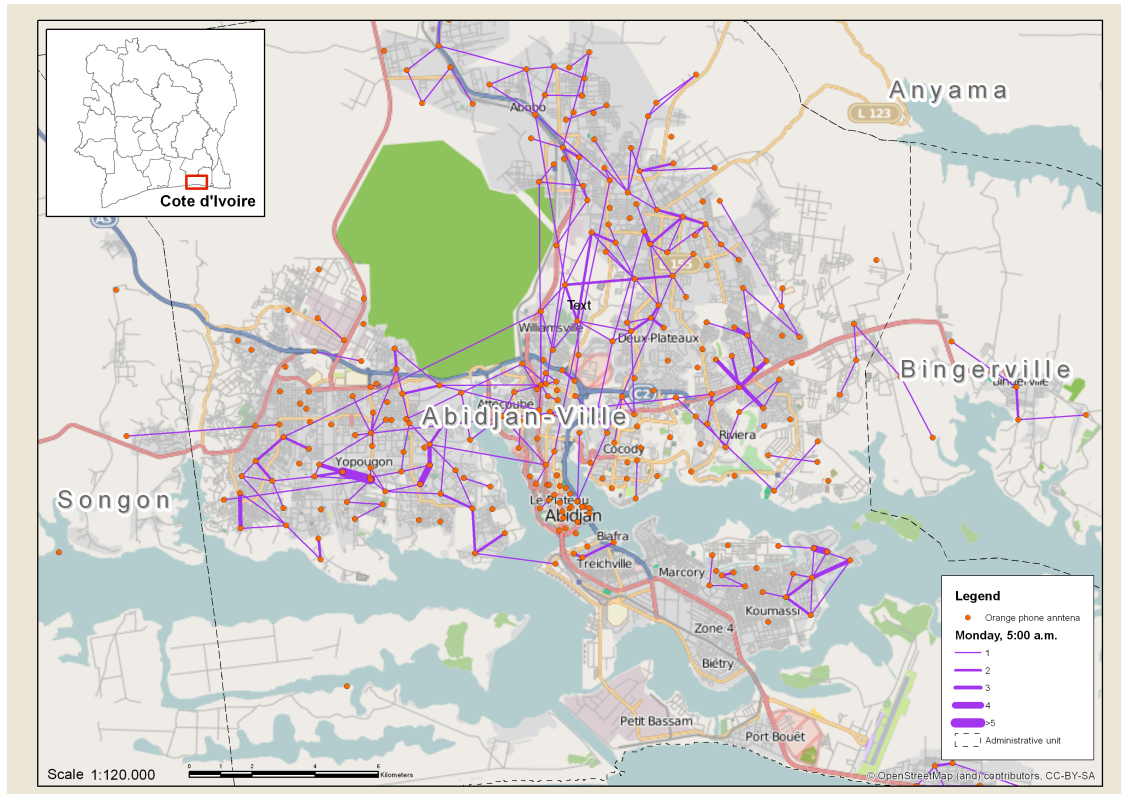

Figure 10: Traffic between antennas

Starting with a file generated from the dataset that contains the geographic coordinates (longitude, latitude) of an antenna origin and those of other target antennas, together with a field that measures the traffic of information between one and the other one ('weight' field), we have generated the corresponding linear entities that join both positions, and the value of the calculated weight has been assigned to them. The visualization is achieved classifying the numerical field (weight) for graduated symbology using an equal interval classification method that divides the range of attribute values into five equal-sized subranges.

Acknowledgements

To Professor Antonio Hernando Esteban[11] (Automatic & systems engineering depto at EUI-UPM) to review the mathematical formalism proposed in this paper.

Also Carmen Vidal, Operations Manager at Paradigma Tecnologico to support this project and the people at Paradigma Labs.

Special thanks to Chris Kostov for his valuable help reviewing this document, carefully spotting grammar, mechanics, and punctuation mistakes, and providing insights into writing in better English.

---

[11]http://www.sia.eui.upm.es/isa/doku.php?id=profesores:ahernando

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.

Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

David P Chrest and William D Wheaton. Using geographic information systems to define and map commuting patterns as inputs to agent-based models. *Methods report (RTI Press)*, 2009(12):906, 2009.

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].

W.J. Craig, T.M. Harris, and D. Weiner. *Community Participation and Geographical Information Systems*. Community Participation and Geographic Information Systems. Taylor & Francis, 2002. ISBN 9780415237529. URL http://books.google.es/books?id=IePxXod45z8C.

W. De Jong, M. Shaw, and N. Stammers. *Global Activism, Global Media*. Pluto Press, 2005. ISBN 9780745321967. URL http://books.google.com.co/books?id=ny6DAAAAMAAJ.

Nathan Eagle, Alex Pentland, and David Lazer. Inferring Friendship Network Structure by Using Mobile Phone Data. *PNAS*, 106(36), 2009.

N.G. Fielding, R.M. Lee, and G. Blank. *The SAGE Handbook of Online Research Methods*. SAGE Publications, 2008. ISBN 9781446206607. URL http://books.google.es/books?id=EeMKURpicCgC.

Carlos Llano Verduras. Localización residencial y movilidad laboral: un análisis del" commuting" de trabajadores nacionales e inmigrantes en la comunidad de madrid. *Cuadernos de economía: Spanish Journal of Economics and Finance*, 29(81):69–100, 2006.

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].

A. Mitchell. *Zeroing In: Geographic Information Systems at Work in the Community*. Esri Press, 1997. ISBN 9781879102507. URL http://books.google.com.co/books?id=Tj6AAAAMAAJ.

D. Moyo. *Dead Aid: Why Aid Is Not Working and How There Is a Better Way for Africa*. Farrar, Straus and Giroux, 2009. ISBN 9780374139568. URL http://books.google.es/books?id=2T_RbtTslzEC.

S. Sassen. *Una sociología de la globalización*. Discusiones. Conocimiento. Katz, 2007. ISBN 9788493543266. URL http://books.google.es/books?id=O8rSxYBetjoC.

A.A. SEN, A.K. Sen, E. Rabasco, and L. Toharia. *Desarrollo y Libertad*. Colección Documento Series. GeoPlaneta, Editorial, S. A., 2000. ISBN 9788408035244. URL http://books.google.com.co/books?id=Jk_bPQAACAAJ.

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/`.

Marius Thériault, Marie-Hélène Vandersmissen, Martin Lee-Gosselin, and Denis Leroux. Modelling commuter trip length and duration within gis: Application to an od survey. *Journal for Geographic Information and Decision Analysis*, 3(1):41–55, 1999.

Fahui Wang. Modeling commuting patterns in chicago in a gis environment: A job accessibility perspective. *The Professional Geographer*, 52(1):120–133, 2000.

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

F Benjamin Zhan, Xuwei Chen, Carol Lewis, Laura L Higgins, and Judy Perkins. Gis models for analyzing intercity commute patterns: A case study of the austin-san antonio corridor in texas. Technical report, Center for Transportation Research, University of Texas at Austin; Springfield, Va., 2008.