Adaptive Non-Parametric Identification of Dense Areas using Cell Phone Records for Urban Analysis

Alberto Rubio[a], Angel Sanchez[b], Enrique Frias-Martinez[a]
[a]Telefonica Research, Distrito C, 28050 Madrid, Spain
[b]Dpto. Ciencias de la Computacion, Universidad Rey Juan Carlos, 28933 Mostoles (Madrid), Spain

Abstract
Pervasive large-scale infrastructures (like GPS, WLAN networks or cell-phone networks) generate large datasets containing human behavior information. One of the applications that can benefit from this data is the study of urban environments. In this context, one of the main problems is the detection of dense areas, *i.e.,* areas with a high density of individuals within a specific geographical region and time period. Nevertheless, the techniques used so far face an important limitation: the definition of dense area is not adaptive and as a result the areas identified are related to a threshold applied over the density of individuals, which usually implies that dense areas are mainly identified in downtowns. In this paper, we propose a novel technique, called AdaptiveDAD, to detect dense areas that adaptively define the concept of density using the infrastructure provided by a cell phone network. We evaluate and validate our approach with a real dataset containing the Call Detail Records (CDR) of fifteen million individuals.

Keywords
Urban analysis, urban dynamics, CDR, dense areas, clustering, baseline correction, mean-shift

1. Introduction
With the increasing capabilities of mobile devices, individuals leave behind traces of their interaction with urban environments. As a result, huge datasets that contain descriptions about urban dynamics are being generated. New research areas, such as urban computing and smart cities, focus on improving the quality of life in an urban environment using datasets obtained from ubiquitous infrastructures, like the examples presented in Liao et al. (2007), Brockmann (2009) and Frias-Martinez et al. (2011).

Traditionally, the study of urban environments has used data obtained from surveys to characterize specific geographical areas or the behavior of groups of individuals. However, new data sources (including GPS, Bluetooth, Wi-Fi hotspots, geo-tagged resources, etc.) are becoming more relevant as traditional techniques face important limitations, mainly the complexity and cost of capturing survey data. One of the new data sources relevant for the study of urban environments are cell phone records, as they contain a wide range of human dynamics information (ranging from mobility to social context and social networks). In this context, cell phones can be considered one of the main sensors of human behavior as they are pervasive in societies and are carried at almost all times.

The process of studying urban environments has the characteristics that machine learning problems have: a dataset containing behavioral information that has to be elicited using machine learning techniques to improve our knowledge of the environment. Nevertheless, due to the size of the data sets being processed (having Terabytes of data is not uncommon), special care should be taken with the architecture needed to process the data and generate the models (Hohwald et al, 2010; Vieira et al., 2010).

One of the key problems for characterizing urban environments is the identification of areas with high density of individuals at specific moments in time. This information is of paramount importance for, among others, urban and transport planners, emergency relief and public health officials, as it provides key insights on where and when there are areas of high density of individuals in an urban environment. Urban planners can use this information to improve the public transport system by identifying dense areas that are not well covered by the current infrastructure, and determine at which specific times the service is in greater demand. On the other hand, public health officials can use the information to identify the geographical areas in which epidemics can spread faster and, thus, prioritize preventive and relief plans accordingly. In this context, cell phone networks offer an ideal data source to detect and characterize urban dense areas.

The problem of dense area detection was initially presented in the data mining community as the identification of the set(s) of regions from spatio-temporal data that satisfy a minimum density value. This problem was initially solved for spatial and multidimensional domains in Wang and Muntz (1997), and later for spatiotemporal domain, as in Ni and Ravishankar (2007), Jensen et al. (2007) and Hadjieleftheriou et al. (2003). In the former proposals, no time dimension was considered, while in the later ones only moving objects, typically represented by GPS sensors that continuously report their locations, are considered. The main limitations of current approaches are: (1) the concept of dense area is defined typically using a threshold that is not set by considering the environment; and (2) all these techniques have a variety of parameters that need to be adjusted, which in the context of urban analysis can be very complex. Ideally, we seek a technique that is non parametric and where the characteristics of the environment are considered to adaptively define the dense areas. This fact is especially important in our context, as the concept of a high density is not the same in downtown and in a suburban area.

The identification of dense areas has also been tackled using Image Processing techniques. With this approach, these urban areas correspond to densely built parts of towns which contain not only buildings but can also include urban streets, railway stations, high-speed roads or malls. Some useful components to be extracted from these urban images are road networks, such as in Hu et al. (2004). Other related problems to dense urban area detection are building extraction such as the work by Xu et al. (2009), or pronounced urban change detection (due to building activities) from aerial images (He and Laptev, 2009). One limitation of these approaches is that dense areas are identified regardless of the actual number of individuals present (which can greatly vary between different time periods) and are just based on concentration of infrastructures. With our approach the identification of dense areas is based on the real number of individuals in these areas at specific moments in time.

In this paper we propose the population-based Adaptive Dense Area Discovery (AdaptiveDAD) algorithm to automatically detect dense areas considering the surroundings. Although our algorithm is going to be designed and tested using data collected from cell phone networks, the same algorithm can be potentially used for any data set generated by other ubiquitous infrastructures, such as location-based services, geo-localized Twitter or geo-localized Flickr.

Note that the focus of this paper is on the detection and study of dense areas, and not in hotspots localization as presented in Agarwal et al. (2006) and Kulldorff (1997). Hotspots, as defined by scan statistics, are the largest discrepancy areas in which an independent variable has statistically different count values from the rest of the geographical areas. Conversely, dense areas, as presented in Crandall et al. (2009), are defined as the (global or local) maxima of the distribution of the function under study. Thus, the information provided by both approaches is different, while hotspots can be used to identify events; dense areas identify regions in space with a minimum critical mass of individuals.

The rest of the paper is organized as follows. After introducing the related work (Section 2), we present the formal problem definition in Section 3. The following Section describes all the pre-processing of data after these are collected from Cell Data Records (CDR) and also shows how to automatically define a grid for the geographical environment under study. Section 5 details the AdaptiveDAD algorithm with its three main stages. After that, we present an experimental evaluation of our method in Section 6. Finally, the conclusions are drawn in Section 7.

2. Related Work

In Geographical Information System (GIS), Urban Planning and Transportation communities, there has been for a long time a variety of models to study urban dynamics. Traditional approaches divide the geographical region under study in zones which exchange population among themselves. Each zone is characterized by a vector of socio-economic indicators, as presented in Benenson (1999), typically collected and generated using surveys. Also, this information can be completed with proxy sources for human mobility such as transport infrastructures (see Brockmann et al. (2006)), or with air connections (see Brockmann (2009)). These approaches provide information about human behavior in a geographical environment but they are very difficult to update and limit the results to a moment in time. In any case, these models substitute humans with derivatives of their activities, ignoring the self-driven nature of human mobility and as such of urban analysis. The use of data originating in pervasive infrastructures captures the self-driven nature of urban environments, complementing traditional approaches. Reades et al. (2007) and Ratti et al. (2006) present initial guidelines on how mobile phone data can be relevant for urban planning and transportation communities.

Previous works on the identification of dense areas have been carried out following three main approaches: (1) density-based clustering techniques; (2) detecting dense fixed-size grids in spatio-temporal data; and (3) spatial-based techniques to detect local maxima areas. Clustering algorithms for spatial, multidimensional and spatio-temporal data have been the focus of a variety of studies such as Ester et al. (1996), Zhang et al. (1996), Kalnis et al. (2005) and Lee et al. (2007). Common to all of the above methods is that clusters with high numbers of objects in a specific geographical area are associated, using spatial properties of the data, to denser regions. Furthermore, all of these methods require choosing some number of clusters or making underlying distributional assumptions of the data, which are not always easy to estimate. There are a variety of solutions for detecting dense areas in spatial domains such as Wang and Muntz (1997), and also in spatio-temporal domains such as Ni and Ravishankar (2007), Jensen et al. (2007) or Hadjieleftheriou et al. (2003). The STING method presented in Wang and Muntz (1997) is a fixed-size grid based approach to generate hierarchical statistical information from spatial data. Hadjieleftheriou et al. (2003) presents another method based on fixed size grids where the main goal is to detect areas with a number of trajectories higher than a predefined threshold. Algorithms using a fixed-size window are proposed by Ni and Ravishankar (2007) and Jensen et al. (2007) to scan the spatial domain in order to find fixed-size dense regions. All of these approaches are specifically designed to work for trajectory data where the exact location and velocity of a trajectory are used in order to aggregate values in each grid for the spatial domain. Unfortunately, these methods cannot be applied to our domain since in mobile phone databases mobile users are not continuously tracked. Some solutions to detect dense areas are based on the identification of local maxima, typically using techniques inherited from Computer Vision (e.g. mean-shift). Mean shift is a non-parametric clustering technique that identifies the modes of a density function given a discrete dataset sampled from that function. Crandall et al. (2009) use mean-shift to identify geographical landmarks from geo-tagged images. In summary, the majority of previous works, among other limitations, typically identify dense areas by using thresholds that need to be stated before hand. In the context of urban analysis, defining this concept is extremely difficult as it depends on the geographical region under study. Not only that, but, within the same geographical region the concept of dense area can not be generically defined by one threshold. In order to effectively identify dense areas for urban applications it becomes necessary to have a non-parametric system that adaptively identifies dense areas considering the population density in the geographical environment.

3. Problem Definition

The adaptive and non-parametric method for identification of dense areas that we propose is based on using the ubiquitous infrastructure provided by a cell phone network. Cell phone networks are built using a set of Base Transceiver Stations (BTS) that are in charge of communicating mobile phone devices with the cell network. A BTS has one or more directional antennas (typically two or three, covering 180 or 120 degrees, respectively) that define a cell and the set of cells of the same BTS define the sector. At any given moment in time, a cell phone is covered by one or more BTSs. Depending on the network traffic, the phone selects the BTS to connect to. The geographical area covered by a sector depends mainly on the power of the individual antennas. Depending on the population density, the area covered by a BTS ranges from less than 1 Km^2 in dense urban areas, to more than 3 Km^2 in rural areas. Each BTS has latitude/longitude attributes that indicate its location and a unique identifier BTS_{id}. For simplicity, we assume that the cell of each BTS is a 2-dimensional non-overlapping region, and we use Voronoi diagrams to define the covering areas of the set of BTS considered. Figure 1((left) presents a set of BTS with the original coverage for each cell, and (right) the simulated coverage obtained using Voronoi diagrams. While simple, this approach gives us a good approximation of the coverage area of each BTS. CDR databases are populated when a mobile phone connected to the network makes/receives a phone call or uses a service in the network (e.g., SMS, MMS, etc). In the process, the information regarding the time and the BTS where the user was located when the call was initiated is logged, which gives an indication of the user's geographical location at a given period in time. Note that no information about the exact user location inside a cell is known. The typical attributes stored in a CDR database include: (1) the originating phone number; (2) the destination encrypted phone number; (3) the type of service (voice, SMS, MMS, etc); (4) the BTS identifier used by the originating number; (6) the timestamp (date/time) of the connection; and (7) the duration of service.

Using the information contained in a CDR database generated from the BTS towers that give coverage to an urban area, we can adaptively identify the dense areas of a city. In order to do so, first the geographical area under study needs to be divided using a grid, where each element will be characterized by the

number of individuals present in that area during the period of time under study. This process requires a transformation from the original data structure (expressed by BTS tower) into the elements of the grid. After that, the three-step AdaptiveDAD algorithm for identification of dense areas is applied. The first step defines a baseline for the geographical area, which is the key element used to adaptively identify dense areas. This baseline is subtracted from the original signal, and, as a result, a set of peaks candidates to be dense areas are identified. The second step applies mean-shift clustering in order to identify which geographical areas are clustered around each one of the peaks identified in the first step. The third (and last) step applies a median filter to each cluster which was previously identified to delimit the contour of each dense area. All steps in the AdaptiveDAD algorithm include a deparametrization phase to produce a more adaptable dense area identification method.

4. Data Preprocessing and Grid Definition

The considered problem of identifying geographical dense areas has the following inputs: (1) the period of time under study, (2) the set of BTSs included in the geographical area considered and (3) the number of unique individuals that have used each BTS during that period of time. The information contained in a CDR dataset can be used to extract the numbers regarding the total amount of unique individuals that have made or received a phone call/SMS/MMS from each BTS using the corresponding filters. As a result, the problem is presented in an anonymized form in which no information about particular users is used.

The characterization of each BTS provides an absolute number which does not consider the actual coverage area of a tower. The first step implies transforming these absolute values into density values by dividing the number of individual users by the covered area of the corresponding tower, in order to have a sense of the density of individuals in the geographical area covered by the BTS. Figure 2(left) presents the absolute number of individuals for each polygon of the Voronoi tessellation and Figure 2(right) shows the same structure after considering the area of each polygon to define the density. It can be observed that while the central part, which contains the downtown of the city, still has very high density values, the density values for the external parts of the city are low, once density is considered due to the large size of the coverage areas of the BTSs. As a result of this process, each element of the Voronoi tessellation has a density of individuals associated with the period of time analyzed.

In order to normalize the data and to avoid the irregularity introduced by the data tessellation, a regular grid is defined using the density values from the tessellation. As such two processes need to be defined: (1) how to transform the density information of the tessellation into the grid selected and (2) how to select the size of the grid. Each element of the grid will have the same value of the Voronoi polygon in which it is included. For those elements of the grid that contain more that one Voronoi polygon, the value will be obtained by weighting the density values by the percentage of the area that it represents, i.e. by approximating the values using a weighted linear interpolation. The regular grid is defined by an increment (measured in degrees) which defines each square of the grid. The value of such increment is highly related with the granularity of the dense areas that we want to identify.

Typically, three levels of dense areas are relevant: urban level, state/regional level and national level. The usual increment values that define the grid are 0.005 degrees for urban levels (approximately 0.5 Km in the equator), 0.1 degrees for state/regional level (approximately 10 Km in the equator) and 0.2 degrees for a country level (approximately 20 Km in the equator). These values have been obtained experimentally. More details of this process can be found in Section 6. Each grid value defines the geographical elements that form dense areas that are relevant at each level. In the extremes, if the geographical area under study is a city, dense areas will be defined by specific blocks of the city and if the area under study is at country level, dense areas will be defined by cities. Figure 3 presents an example of the effect of the grid size in the definition of dense areas. The figure on the left represents a city, in which the grid has been defined with a 0.005 degree increment. As a result different parts of the city can be identified as dense areas. The figure on the right represents the same geographical area but with a grid defined with 0.1 degrees increment, better suited for regional levels. As a result, the whole city is transformed in just one peak and only one dense area will be identified.

5. AdaptiveDAD: Adaptive non-parametric Dense Area Detection algorithm

The proposed algorithm has been designed for the adaptive identification of dense areas in a geographical environment. The algorithm consists of three main stages: baseline extraction, mean-shift clustering and independent filtering the contour of each dense area. The baseline extraction aims to quantify the geographical dense areas as a function of their respective environments; the mean-shift clustering is applied to identify the geographical extensions of each dense area; and the filtering stage removes some remaining noise and it also delimitates accurately the identified areas.

5.1. Baseline Extraction

The goal of this stage is to find a surface that fits to a three-dimensional (3D) function in such a way that the local maxima are isolated for further study. This process, known as baseline suppression, will allow for fair comparisons between these maxima, since they will not be influenced by their respective environments. The baseline noise always blurs signals and spoils analytical results, especially in multivariate analysis. Consequently, it is first necessary to compute the baseline and remove its effect on the signal to perform further data analysis. This baseline filters the original signal by preserving its peaks and its shape, as it is shown in Fig. 4, where the 2D original signal (in blue) is transformed into the filtered signal (in black) after subtracting the estimated baseline (in red).

Several polynomial fitting methods to compute the baseline of a signal have been proposed (Andrade and Manolakos, 2003). To remove the baseline effect, we have used a recent algorithm called airPLS (adaptive iteratively reweighted Penalized Least Squares) proposed by Zhang et al. (2010). This is an unsupervised method which does not need from any previous peak detection. It works by iteratively modifying the SSE (Sum of Square Errors) weights between the computed baseline and the original signal. The weights of the SSE are obtained adaptively using the difference between the previously fitted baseline and the initial signal. This baseline extraction method is accurate, flexible and also computationally efficient. Although the method has been mainly applied to chromatographic 2D signals, it can be easily adapted for our 3D signals representing dense urban areas.

For the baseline correction purpose, different penalized least squares algorithms have been proposed (in the analytical chemistry domain, mainly). These algorithms balance the fidelity of the original signal and the roughness of the adjusted signal. In this context, Eilers and Boelers (Eilers and Boelers, 2005) presented an asymmetric weighted least squares (smoothing) method for baseline correction. Given a 2D signal of length m that is sampled at equal intervals, the goal is to iteratively minimize the following functional S which expresses a trade-off between the similarity of the original signal y and the adjusted one z, and the smoothness degree of z

$$\begin{aligned} S &= \sum_{i}^{m}(y_i - z_i)^2 + \lambda\sum_{i}^{m}(\Delta^2 z_i)^2 \\ \Delta^2 z_i &= (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2} \end{aligned} \quad (1)$$

where the parameter λ sets the influence of each of the two terms in S. The first term evaluates the similarity between the signals y and z, while the second term penalizes the non-smooth behavior of z (i.e. this penalization is attenuated when signal z becomes smoother). Next, equation (1) is generalized with the introduction of a vector of weights w (in the first term of S) to have a better fitting between the original and the adjusted signal:

$$S = \sum_{i}^{m} w_i(y_i - z_i)^2 + \lambda\sum_{i}^{m}(\Delta^2 z_i)^2 \quad (2)$$

In this formulation, a parameter is p (representing the asymmetry of the least squares weights) is used to compute the weights w_i as follows: $w_i=p$ if $y_i>z_i$, and $w_i=1-p$ otherwise. The minimization of S produces the following system of equations:

$$(W + \lambda D' D)z = Wy \quad (3)$$

where: W=diag(w) and D is a matrix of differences: $Dz = \Delta^2 z$. This is a large and very sparse system with m equations where only the main diagonal and the two above and sub-diagonals are different from zero. This system can be solved iteratively (the weights vector is initialized to one and updated during the

iterations) to determine the adjusted signal z. According to the value of parameter λ, the baseline function can be applied to different purposes. For example, a light smoothing will remove noise, while a strong smoothing gives the slowly varying trend of a signal (Eilers and Boelers, 2005). However, the method described has as its main drawback the need to optimize the value of the asymmetry parameter p to achieve satisfactory baseline correction results. The airPLS algorithm by Zhang et al. (2010) does not require the inclusion of a parameter p and computes the weight vector w iteratively. The initial value w^0 is set to 1, and at iteration t the weights w are computed as follows:

$$w_i^t = \begin{cases} 0, & y_i \geq z_i^{t-1} \\ e^{\frac{t(y_i - z_i^{t-1})}{|d^t|}}, & y_i < z_i^{t-1} \end{cases} \quad (4)$$

where d^t is a vector of negative elements of the differences between y and z^{t-1} in the iteration t. The iterative process stops after a maximum number of iterations (*maxIter*) or when the condition: $|d_t| < 0.001 \times |y|$ is reached. In the airPLS algorithm, the peak points are gradually removed and the baseline points are preserved in the vector w, as shown in Zhang et al. (2010).

The airPLS algorithm has three input parameters: λ that represents the balance between error and smoothness terms in S, *maxIter* that represents the maximum number of iterations of the algorithm, and O is the order of the polynomial z used to adjust to the input signal y. To de-parameterize the algorithm, the polynomial-order parameter O was first set in our approach, since there are few possible values to test and it also this parameter conditions the values of the other two ones. We chose $O = 2$ as a good trade-off between fidelity to the original signal and computational efficiency. After that, a two-level resolution search strategy (the first one at a rough level with a λ step of 0.5 and the second one at a finer level with a λ step of 0.05) was used to estimate good combinations of (*maxIter*, λ)-parameters for our problem. This method produced very good results for the signals describing dense areas that we have used (although it does not always guarantee to obtain the optimal values of parameters). Best results were achieved in our experiments for *maxIter* = 6 (this value is in keeping with a work by Eilers and Boelers (2005) who considered: $5 \leq maxIter \leq 10$ as an appropriate interval of iterations in the practice), and the following interval of values for parameter λ: $0.35 \leq \lambda \leq 4.5$.

Having computed the baseline for 2D signals, it is necessary to adapt this task to a 3D environment as it is the case for the data used to detect dense geographic areas. The proposed method consists of first applying the 2D baseline algorithm to each row of the original data, then to each column of the data, and finally to combine the results obtained for each one of the rows and each one of the columns, thus obtaining a 3D representation of the dense areas. Three different ways of combining the rows and columns results were tested: (1) "average value" of each point at both corresponding row and column 2D baseline results; (2) "minimum value" of corresponding row and column 2D baseline results and (3) "maximum value" of corresponding row and column 2D baseline results. We adopted the "maximum value" model since it produced for our experiments the smallest adjustment error with respect to the original data and it also preserved best the shape of these data. Fig. 5 illustrates with an example the three stages of the 3D baseline modeling process.

5.2. Mean-Shift Clustering

The next stage of the method aims to filter some noise produced by the application of the baseline algorithm and, after that, to identify the correct dense areas from the filtered data (i.e. to delimit properly the environment of each dense area identified in the baseline extraction stage). As the baseline algorithm can produce small peaks with associated environments which do not correspond to dense areas, we compute the local maxima from original data and compare the positions of these maxima with those ones corresponding to local maxima obtained after the baseline computation. Those local maxima whose positions do not approximately correspond on both models are removed and considered as noise. This is carried out by considering a reduced neighborhood around each local maxima and using mathematical morphology operators.

Next, we search for a clustering algorithm to determine the environments of dense areas from the filtered local maxima. This clustering algorithm should make it possible that the computed clusters have any arbitrary shapes. Moreover, the algorithm must be de-parameterized, tolerant to noise and also flexible with respect to the size of datasets.

According to the clusters shape restriction, some data clustering algorithms like *k*-means and BIRCH are discarded (Jain et al. (1999)). Moreover, the presence of parameters in the algorithm also rules out other methods like DBSCAN, WaveCluster or DENCLUE. After analyzing a collection of clustering methods for the proposed task, we conclude that the algorithm which best fits all previous requirements is mean-shift. This method was originally proposed by Fukunaga and Hostetler (1975) as a non-parametric clustering algorithm oriented towards the analysis of feature spaces following complex multimodal distributions. The goal is to find on these spaces a set of clusters with arbitrary shape without any prior knowledge. Mean-shift has been mainly applied to image analysis tasks like smoothing filtering and contour-based segmentation (Comaniciu and Meer (2002)).

Mean-shift is a non-parametric iterative algorithm for density gradient estimation using a generalized kernel approach. This algorithm considers the points of the feature space as a probability density function (pdf) $f(x)$. Dense regions in the feature space correspond to local maxima (or modes) of the pdf. The goal is to locate these modes given a set of n discrete points of it. To achieve this, mean-shift performs the gradient ascent on each of the points on the local estimated density function until convergence.

Given a kernel function K with radius parameter h, and a set of n data points $x_i \in R^d \ (\mathrm{i}=1..n)$, the kernel density estimator for these points is:

$$f(x) = \frac{1}{nh^d} \sum_{i}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5}$$

Tipically, the kernel $K(x)$ is a radial symmetric function of $\|x\|^2$, such as: $K(x)=c\,k(\|x\|)^2$, where: c is a strictly positive normalization constant (that makes $K(x)$ integrate to one), and $k(x)$ is a non-negative, non-increasing and piecewise continuous function called profile of the kernel (that is defined only for $x \geq 0$). Local maxima of the density function (and the other stationary points) correspond to zeros of the gradient function: $\nabla f(x)=0$. Then, computing the gradient of the density estimator in eq.(5), and after some algebraic manipulations, we obtain:

$$\nabla f(x) = \underbrace{\frac{2c}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \right]}_{term1} \underbrace{\left[\frac{\sum_{i=1}^{n} x_i g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)} - x \right]}_{term2} \tag{6}$$

where: $g(x)=-k'(x)$ (i.e. derivative of selected kernel profile). In eq. (6) the *term1* is proportional to the density estimate at x, while *term2* is called the mean shift vector $m_h(x)$:

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)} - x \tag{7}$$

which points towards the direction with maximum increment in density. The mean-shift algorithm is an iterative procedure that performs the following steps when applied at each iteration t to each point x_i^t of the dataset (where $i=1..n$):

(a) Compute the mean-shift vector for the point: $m_h(x_i^t)$, according to eq. (7).
(b) Perform gradient ascent (i.e. update point position): $x_i^{t+1} = x_i^t + m_h(x_i^t)$.
(c) Repeat steps (a) and (b) until convergence (i.e. where the gradient of density function at the point is zero: $\nabla f(x_i)=0$).

A proof of convergence of the mean-shift algorithm can be found in Comaniciu and Meer (2002).

When mean-shift clustering is applied to our urban dense areas estimation problem, the input is a 3D discrete grid space where the height of each point represents its density. First, using a random seed point and a neighborhood size δ, the mean-shift procedure is applied until a local maximum (or mode) of the

pdf is found. All the visited points to find these maxima are marked. The set of all points that converge to the same maxima (or mode) of the pdf are associated with the same cluster. This process is next repeated by choosing a non-visited point to start again as a seed to find new local maxima of the pdf. The algorithm finishes when all the 3D grid positions have been visited or when a stated number of iterations are reached. This is a non-deterministic method since its result will depend on the starting (seeds) positions. Fig. 6 presents an example of the result produced by mean-shift when it is applied to the considered problem.

If the bandwidth δ which defines the neighborhood radius for each point is very small, the convergence to a local maximum becomes difficult; otherwise, when this value is big, several local maxima can be associated with the same cluster. Consequently, a correct bandwidth value should serve to ensure that each 3D domain point is efficiently associated with one cluster which corresponds to every significant local maximum in the considered data space. To estimate the bandwidth parameter δ, we first extract all the local maxima of the analyzed geographical data and then compute the Euclidean distance between the closest maxima. This minimal distance defines the value of δ to be chosen.

5.3. Filtering of Dense Areas

The last stage of AdaptiveDAD aims to filter some remaining impulsive noise and also to delimitate the regions corresponding to each dense area previously obtained. To reduce the computational complexity of this task, this filtering was not globally applied to the complete 3D geographical mesh but separately on each of the detected clusters (i.e. the main signal peaks and their associated regions) obtained after the use of mean-shift algorithm. The proposed filtering of dense areas consists of two successive stages: (a) removal of the remaining impulsive noise and (b) smoothing of the surface regions while preserving the shape of the peaks. Our goal is first to remove some impulsive noise still present, and also to smooth the 3D landscape regions while preserving the shape of the peaks.

To remove the impulsive noise, a simple and effective non-linear 3×3 median filtering (Perreault and Hebert (2007)) was locally applied only on the isolated impulse peaks detected. Figure 7(a) shows a zoomed region of the 3D geographic area (i.e. at a higher grid resolution) where some small peaks are present (this zoomed area is enclosed by a black rectangle). In Figure 7(b) these impulse noise peaks have been removed in the zoomed region after the median filtering application. After that, we apply a second filtering stage, using a "shape filtering" algorithm that identifies and extracts the points of the 3D grid which belong to the peaks and their environments and removes background noise. For this stage, we also propose a de-parameterized solution with two phases: (1) computing a gradient measure of each point and (2) shape filtering based on the previous measure. Finally, we de-parameterize the solution and evaluate it using a fitness function.

The pseudo-code presented in Figure 8 computes the gradient measure on each 3D grid point (where *card* is the cardinal of a set and $|\bullet|$ represents the absolute value function). Using the 3D data grid m (after median filtering) and the structure containing the gradient values *filt_m* and a threshold γ ($0 \leq \gamma \leq 1$), the "shape filtering" is applied separately on all peaks of the 3D geographic mesh. Figure 7 shows an example of this second filtering applied to dense geographical areas meshes. To evaluate the quality of the final filtering result, we use a fitness function that considers two goals: maximize the average height of the remaining peak points (after this last filtering) and maximize the number of points of the original grid that remain after filtering. In our experiments, we observed that giving the same weight of 0.5 to both goals produced good results when evaluating the obtained solutions. However, these weights can be tuned depending on the final application of the framework.

The proposed "shape filtering" used two parameters: the neighborhood radius r required for computing the considered local gradient measure (see pseudo-code of Figure 8) and the threshold γ used to decide if a 3D grid point still remains after the filtering (i.e. when its gradient value is greater than γ). As the ranges of parameters considered were reduced, we tested all possible combinations of (r, γ), where: $1 \leq r \leq 5$ and $0.1 \leq \gamma \leq 0.9$ (using increments of 0.1). Each computed solution was evaluated according to the proposed fitness function and the best solution is chosen as the result of the "shape filtering" process. Figure 7(b) presents the result of this smoothing filter on the range surface of Figure 7(a).

6. Experimental Evaluation

In order to evaluate and validate AdaptiveDAD, we collected aggregated cell phone data from Mexico for a period of four months considering 15 million users that generated over 300 million interactions. The set

of users represent to a large extent the variety of individual characteristics and socio-economic factors of the country. In order to study the dynamics of dense areas we considered six different time slots: (1) 6am to 9:59:59am; (2) 10am to 1:59:59pm; (3) 2pm to 5:59:59pm; (4) 6pm to 9:59:59pm and (5) 10pm to 1:59:59am and (6) 2am to 5:59:59am. We will focus our studies mainly in the first five, as the last time slot registers a reduced mobility and/or cell phone activity.

The data was aggregated using a set of tuples: (BTS_{id}, *number of different users*), for each time slot considered, so no personal information was actually gathered. Previous studies have highlighted human behavioral differences between weekdays and weekends (Frias-Martinez et al. (2003), Candia et al. (2008), Soto et al. (2011)), and such differences should be reflected in the use of the BTS towers. Two sets of tuples were generated: one considering weekdays (from Mondays through Fridays) and one for weekends (Saturdays and Sundays). The *number of different users* was calculated as the average of different individuals that made or received a phone call/SMS/MMS during weekdays or weekend days in the time slot considered. Public holidays (4 days in total) were filtered from the data set to avoid considering irregular behaviors. In order to evaluate and validate our algorithm, the following subsections present the application of AdaptiveDAD at both urban level and national levels and also in a study focused on city dynamics.

6.1. Experimental Evaluation: Urban Level

In the experimental evaluation we are going to detail the intermediate processes of AdaptiveDAD focusing on an urban environment. For this purpose, we have chosen the metropolitan area of Guadalajara (state of Jalisco, Mexico), that covers over 400 Km^2 and has over 4,500,000 inhabitants. Figure 9 shows the geographical area under study and some of the main landmarks of the city, including the subway system. The two subway lines in the city run East-West (L1) and North-South (L2) with one central station in common. For reference purposes, the central station is denoted by C. The downtown area is geographically located around C, E1, E2, E3 and E4 stations where we can find university buildings, government offices, markets and commercial streets. The rest of L1 services mainly correspond to residential areas. Regarding L2, around S3 to S11, there are mainly residential neighborhoods with light industrial areas. Stations N2 to N7 serve residential areas and some malls. For the areas not covered by the metro system, as a general rule, there is a mixture of residential areas (with different densities) and light industrial areas (with the north and especially north-west having more affluent areas than the south). As an experimental validation, it is expected that, when identifying dense areas for the city of Guadalajara, the highest area would be around downtown as it is here where the commercial and institutional part of the city is located.

Figure 10 graphically details the main steps of AdaptiveDAD when applied to the metropolitan area of Guadalajara considering weekdays and the time slot from 6am to 9:59:59am with a grid step of 0.005 degrees. Figure 10(a) presents the data after the grid has been defined and the information has been transformed from BTS tower into the elements of the grid. It can be observed the differences between the city and the metropolitan area, and within the city, downtown shows a high peak. This graph highlights the limitations that non-adaptive dense area detections techniques suffer, because the areas identified with those techniques will be focused around the set of absolute highest peaks, without considering the relevance of that peak within its environment. As a result, local dense areas are lost in the process. Figure 10(b) shows the baseline constructed with *airPLS* (note that the scale is not the same as in (a)). After subtracting the baseline from Figure 10(a), the result is presented in Figure 10(c). As a result the candidate dense areas are better defined, as all the noise is eliminated from the original 3D signal. Figure 10(d) presents the final result after applying mean-shift and the median filter. Figure 11 shows the dense areas identified by AdaptiveDAD and presented in Figure 10(d) over a map of Guadalajara. As expected, the main peak corresponds to downtown and it is located over the main commercial part of the city.

In order to highlight the qualitative advantages of AdaptiveDAD, Figure 12(a) presents the dense areas identified for weekdays in the 6am to 9:59:59 time slot when using a traditional dense area algorithm that we implemented based on the concept of density and filtering dense areas above a threshold. In this case the colors indicate the different dense areas and the warmer the color the higher the density. Figure 12(b) presents the dense areas identified in the same context by AdaptiveDAD. As it can be noticed, the adaptive nature of AdaptiveDAD identifies the parts that are above the average of the surroundings, which avoids that the whole downtown being identified as just one dense area, even though, their density is higher than any other part of the city. The algorithm also identifies the parts that are above the average of the surroundings, which allows for a much finer identification of dense areas. The second highest peak of figure 12(b), on the left, corresponds to the biggest mall of the city, which during the time slot

considered, seems to attract a high number of individuals when compared to its surroundings. This is a good example of a dense area, that is not identified when using traditional techniques, as the intensity surrounding the peak and its proximity to downtown hides any other relevant local dense areas.

6.2. Experimental Evaluation: National Level

Running AdaptiveDAD at a national level is also an easy way of validating the algorithm as the dense areas identified should locate the main metropolitan areas of Mexico. Figure 13 presents the dense areas identified for the whole country during weekdays for the time slot 10am to 1:59:59pm with an incremental grid of 0.2 degrees. As expected the main metropolitan areas of Mexico (Mexico City, Guadalajara and Monterrey with a population of 20 million, 4.5 million and 4 million, respectively) are identified as the top three dense areas. Some resort towns of the Pacific coast are clearly identified, such as Acapulco, Puerto Vallarta, Mazatlan and Culiacan. In this particular case, there is no relevant difference between the considered time slots analyzed.

6.3. City Dynamics

One of the main limitations of traditional urban analysis techniques is that data are usually aggregated over a period of time that just captures a moment in time. As a result, it is difficult to study the evolution over time of relevant urban analysis elements, such as dense areas. Nevertheless the use of pervasive infrastructures makes it possible to follow and study the evolution of urban dynamics over time. In this section, we present the evolution over time of the dense areas for the city of Guadalajara considering the five time slots defined previously. It has to be noted that the dynamics reflected by the evolution of dense areas does not necessarily imply flocks of individuals moving from one area to another, but just density of individuals changing over time.

Figure 14 shows the evolution of dense areas for the five time slots defined during weekdays, and the result gives an indication of how people moves in the city. It is important to remember that with our algorithm the areas identified are characterized by a density of individuals that is higher that in its surroundings. This means that areas not identified as dense areas can have a high density although there is not a change when compared with its surroundings.

Following the temporal sequence, it can be observed that in general the dense areas remain constant in space although their relevance changes over time. Starting in the 6am to 9:59:59am slot, we can observe that the main dense area appears around C, E1 and E2 metro stations, around the commercial, governmental and commercial district. The second dense area is identified west of C, where the main malls of the city are located. Nevertheless, in this time slot the relevance of these dense areas is limited (i.e. these dense areas identified are not different from their surroundings). The evolution over the next three time slots, (10am to 1:59:59pm, 2pm to 5:59:59pm and 6pm to 9:59:59pm, respectively) highlights the dense area in downtown over the rest of the areas identified, although the dense area around the malls also increases in relevance. The last time slot (10pm to 2am), returns to a situation very similar to the first time slot although in this case the relevance of the main two dense areas is very similar. The rest of dense areas identified in all time slots are focused in residential neighborhoods. The identification of dense areas in residential neighborhoods is tightly related to the density of housing, where lower income neighborhoods tend to have a higher density than more affluent ones. This is probably one of the reasons why residential dense areas are identified mainly in the south and east and less in the northwest. In any case, the relevance of all these areas is small compared to the two main dense areas (downtown and malls).

The same sequence was obtained for weekends, and a very similar set of dynamics was observed. The main difference was that the two main dense areas around C metro station and the main mall of the city were also present but with a reduced intensity when compared to weekdays.

A direct application of this knowledge regarding the social dynamics of the city is to help in the decision process of the design of the public transport infrastructure. In general, it can be observed that there is an alignment between the metro infrastructure and the dense areas identified in the center and east of the city. Nevertheless, the dense areas located west of C metro station are not covered by the metro lines. From an urban planning perspective, this information could be used to propose line extensions to public transport officials.

7. Conclusions

Ubiquitous computing infrastructures are opening new doors for the study of social dynamics, especially in the field of urban planning and transportation design. While traditional approaches are based on questionnaires, which implies cost and time limitations, our approach, based on using data captured over pervasive networks, overcomes these problems and brings new advantages, like the ability of focusing the studies in particular social group (elders, tourists, socio-economic levels, etc). Our approach is not intended to substitute traditional urban analysis approaches but to complement and improve them.

The identification of dense areas is a key topic for a variety of social dynamic studies, and as such, has received attention from a variety of research fields. Nevertheless, the techniques used so far have some limitations, mainly the lack of adaptability to the locality of the information. In this paper, we have proposed the AdaptiveDAD algorithm to identify dense areas from Call Detail Records (CDR). Our algorithm overcomes the limitations of traditional techniques by adapting to the local density of data and as a result identifies dense areas that, using traditional approaches, would have been ignored. We tested and validated AdaptiveDAD using a real CDR dataset of over 300 million interactions, showing how we can characterize the dynamics of an urban environment from a dense area perspective.

As future work we plan to combine our technique with data originating from other urban sensors, such as traffic, and analyze the applicability of our work for the recommendation of urban planning decisions.

Acknowledgements

This research has been partially supported by the Spanish project TIN2008-06890-C02-02.

References

Agarwal, D., McGregor, A., Phillips, J., Venkatasubramanian, S., Zhu, Z., 2006. Spatial scan statistics: Approximations and performance study. In: 12th ACM SIGKDD Conference, 24-33.

Andrade, L., Manolakos, E., 2003. Signal background estimation and baseline correction algorithms for accurate DNA sequencing. Journal of VLSI Signal Processing Systems 3 (35), 229–243.

Benenson, I., 1999. Modeling population dynamics in the city: from a regional to a multi-agent approach. Discrete Dynamics in Nat. and Soc. 3 (2-3), 149-170.

Brockmann, D., 2009. Human mobility and spatial disease dynamics. Review of Nonlinear Dynamics and Complexity - Wiley.

Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439, 462-465.

Candia, J., Gonzalez, M., Wans, P., Schoenharl, T., Barabasi, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A: Math. Theor. Vol. 41.

Comaniciu, D., Meer, P., 2002. Mean-shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5), 603–619.

Crandall, D. J., Backstrom, L., Huttenlocher, D., Kleinberg, J., 2009. Mapping the world's photos. In: 18th WWW Conference, 761-770.

Eilers, P., Boelers, H., 2005. Baseline correction wih asymmetric least squares smoothing. Technical Report, Leiden University Medical Centre, The Netherlands.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A local-density based spatial clustering algorithm with noise. In: ACM SIGKDD.

Frias-Martinez, E., Williamson, G., Frias-Martinez, V. 2011. An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information. In: The 3rd IEEE Int. Conf. on Social Computing (SocialCom 2011), Boston, MA

Frias-Martinez, E., Karamcheti, V. 2003. A customizable behavior model for temporal prediction of web user sequences. In: WEBKDD 2002-Mining Web Data for Discovering Usage Patterns and Profiles, 66-85.

Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function with applications in pattern recognition. IEEE Transactions on Information Theory 21 (1), 32–40.

Hadjieleftheriou, M., Kollios, G., Gunopulos, D., Tsotras, V., 2003. On-line discovery of dense areas in spatio-temporal databases. Advances in Spatial and Temporal Databases, 306–324.

He, L., Laptev, I., 2009. Robust change detection in dense areas via SVM classifier. In: Proc. GRSS/ISPRS Workshop on Data Fusion and Remote Sensing over Urban Areas (URBAN), 20–22.

Hohwald, H., Frias-Martinez, E., Oliver, N., 2010. ARBUD: A reusable architecture for building user models from massive datasets. In: Workshop on Pervasive User Modeling and Personalization (PUMP).

Hu, X., Tao, C.V., Hu, Y., 2004. Automatic road extraction from dense urban area by integrated processing of high resolution imagery and LIDAR data. In: Proc. IAPRSIS XXXV-B3, 288–292.

Jain, A.K., Murty, N. M., Flynn, P.J., 1999. Data clustering: A review. ACM Computer Surveys 31 (3), 264–323.

Jensen, C., Lin, D., Ooi, B., 2007. Continuous clustering of moving objects. IEEE Trans. Knowl. Data Eng. 19 (9), 1161-1174.

Kalnis, P., Mamoulis, N., Bakiras, S., 2005. On discovering moving clusters in spatio-temporal data. In: SSTD. pp. 364–381.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics-Theory and methods 26 (6), 1481-1496.

Lee, J.-G., Han, J., Whang, K.-Y., 2007. Trajectory clustering: a partition-and-group framework. In: SIGMOD Conf., 593-604.

Liao, L., Patterson, D. J., Fox, D., Kautz, H., 2007. Learning and inferring transportation routines. Artificial Intelligence 171 (5-6), 311-331.

Ni, J., Ravishankar, C., 2007. Pointwise-dense region queries in spatio-temporal databases. In: 23rd IEEE Int. Conf. Data Engineering, 1066-1075.

Perreault, S., P. Hebert, 2007. Median filtering in constant time. IEEE Transactions on Image Processing 16 (9), 2389–2394.

Ratti, C., Williams, S., Frenchman, D., Pulselli, R. M., 2006. Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design 33(5), 727-235.

Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C., 2007. Cellular census: Explorations in urban data collection. Pervasive Computing, 30-38.

Soto, V., Frias-Martinez, E. 2011. Automated Land Use Identification using Call Detail Records In: 3rd ACM Int. Workshop on Hot Topics in Planet-Scale Measurement, in conjuntion with ACM MobiSys2011**,** Washington DC, 2011.

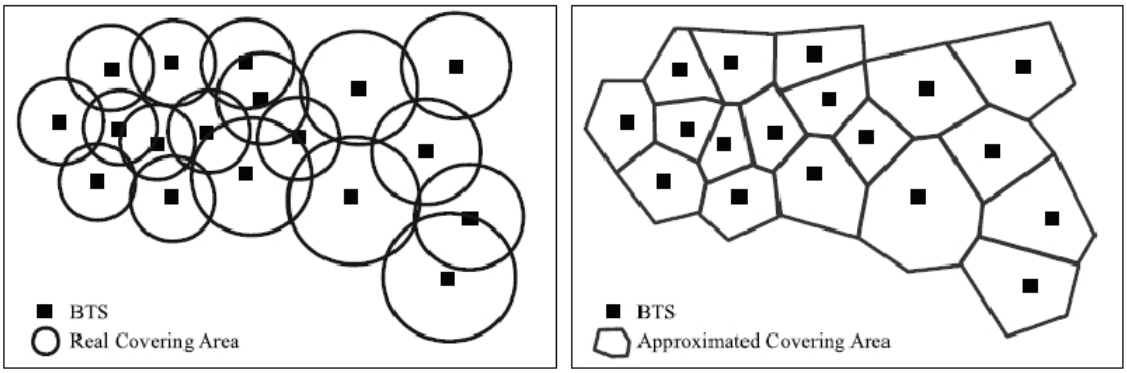
Vieira, M.R., Frias-Martinez, E. Bakalov, P., Frias-Martinez, V. 2010. Querying spatio-temporal patterns in mobile phone-call databases. In: 11th Int. Conf. Mobile Data Management (MDM), 239-248.

Wang, W., Muntz, R., 1997. STING: A statistical information grid approach to spatial data mining. In: Proc. Int. Conf. VLDB, 186-195.

Xu, G., Gao, L., Yuan, X., 2009. Building extraction from aerial imagery based on the principle of confrontation and priori knowledge. In: Proc. Second International Conference on Computer and Electrical Engineering, 363–367.

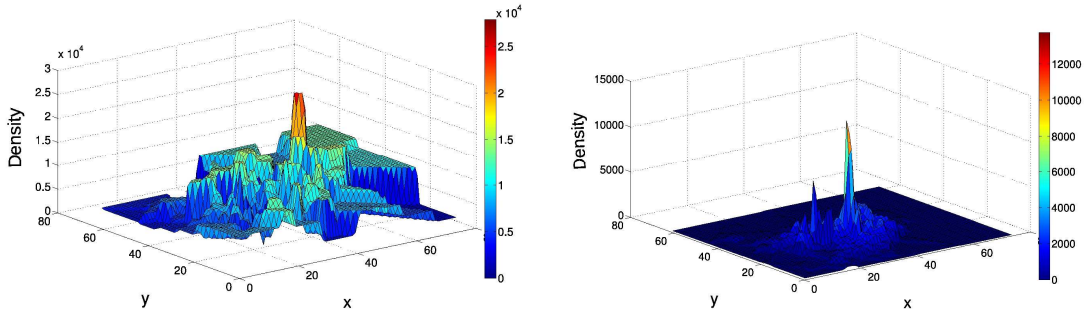
Zhang, Z.-M., Chen, S., Liang, Y.-Z., 2010. Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst 135 (5), 1138–1146.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: An efficient data clustering method for very large databases. ACM SIGMOD Record 25 (2), 103-114.



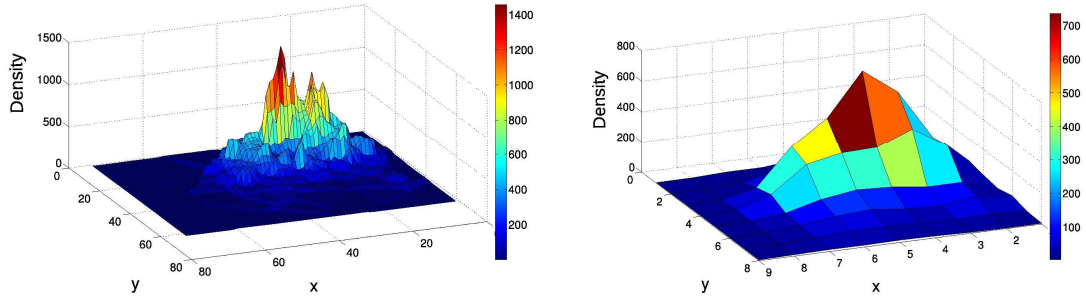
(left) (right)

Figure 1: (left) Example of a set of BTS and their coverage and (right) approximated coverage obtained applying Voronoi tessellation.



(left) (right)

Figure 2: (left) Absolute signal obtained for an urban area where the peak corresponds to downtown and (right) the same urban area representing densities instead of absolute value for each element of the tessellation.



(left) (right)

Figure 3: (Left) Urban geographical area with a 0.005 grid definition and (right) with a 0.1 degree definition.

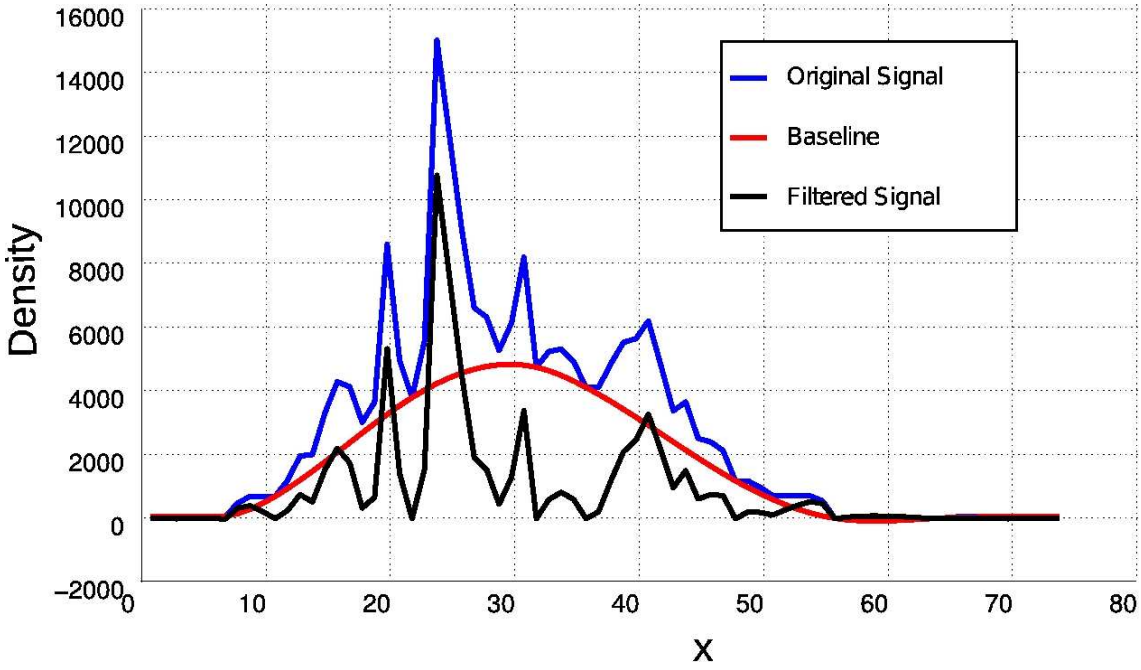
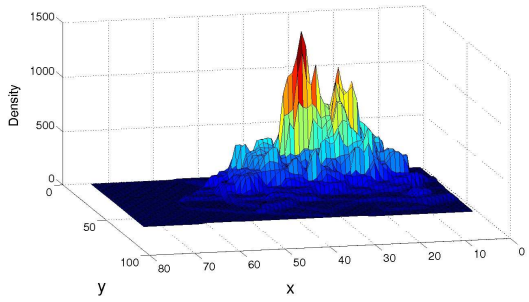
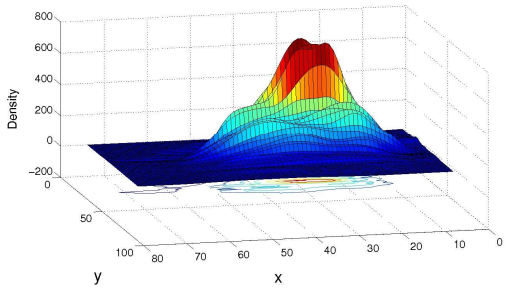


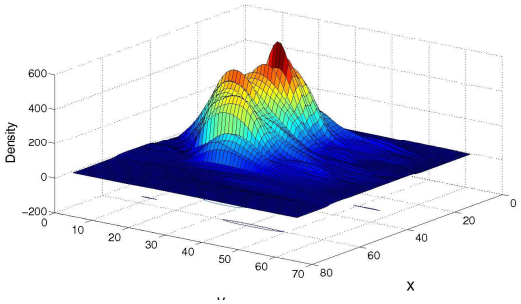
Figure 4: Example of baseline estimation for a 2D signal.



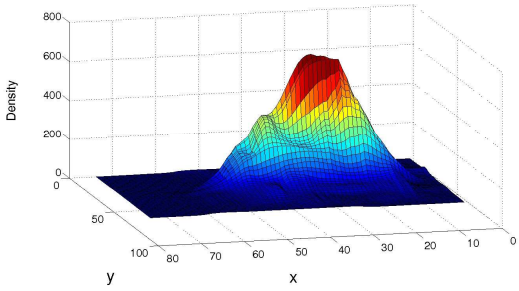
(a) Original 3D data.



(b) Baseline application by rows.

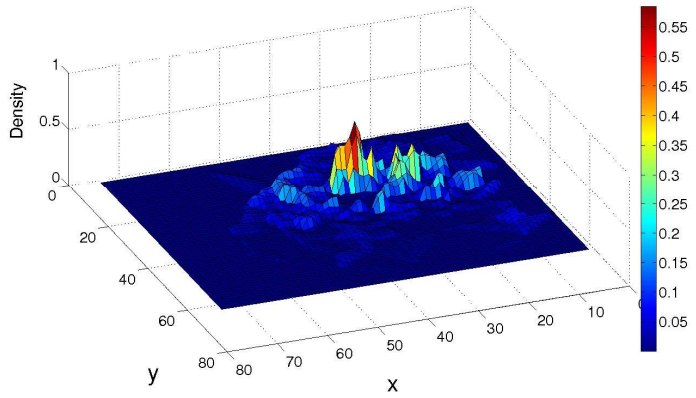


(c) Baseline application by columns.

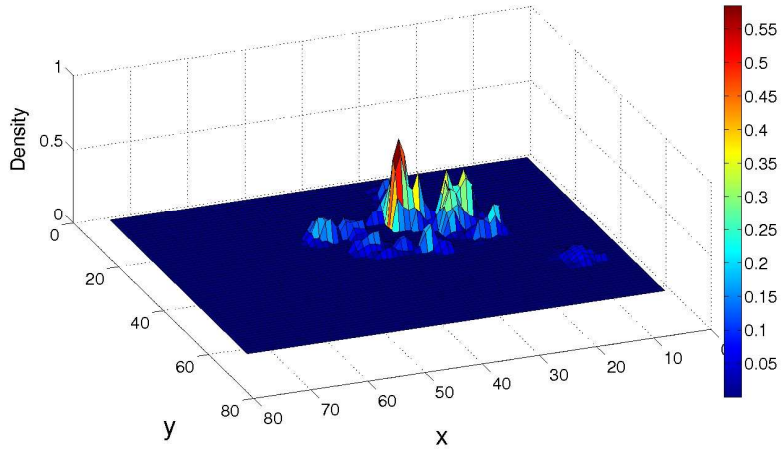


(d) Final filtered 3D data.

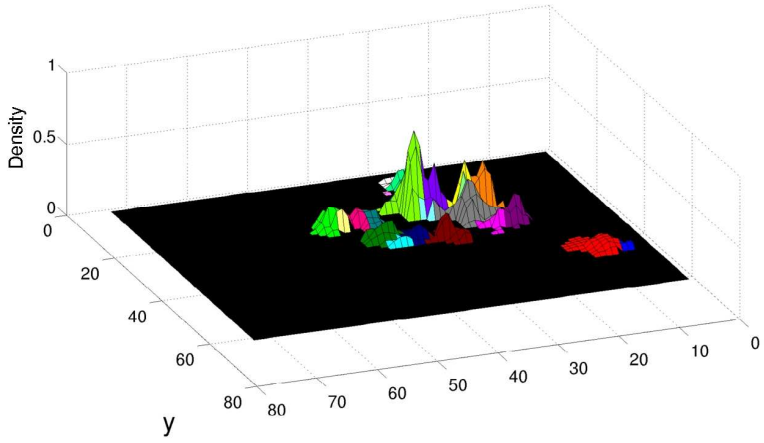
Figure 5: Example of 3D baseline modeling process.



(a) Original 3D data.

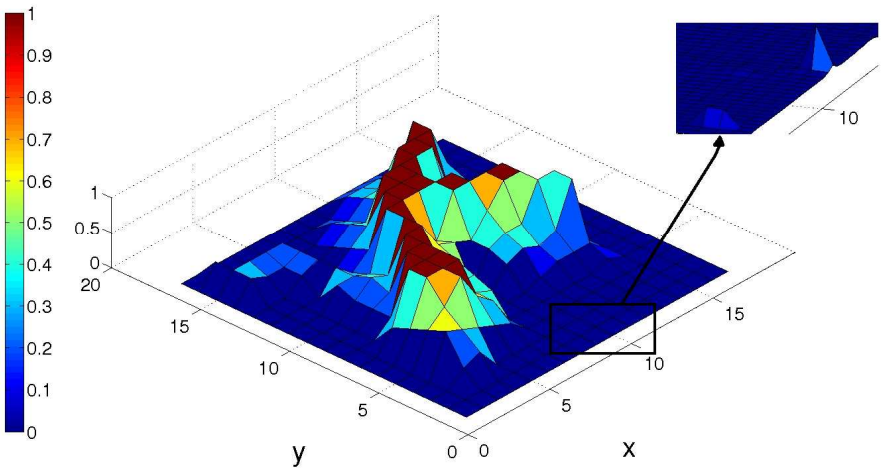


(b) Filtering pre-processing.

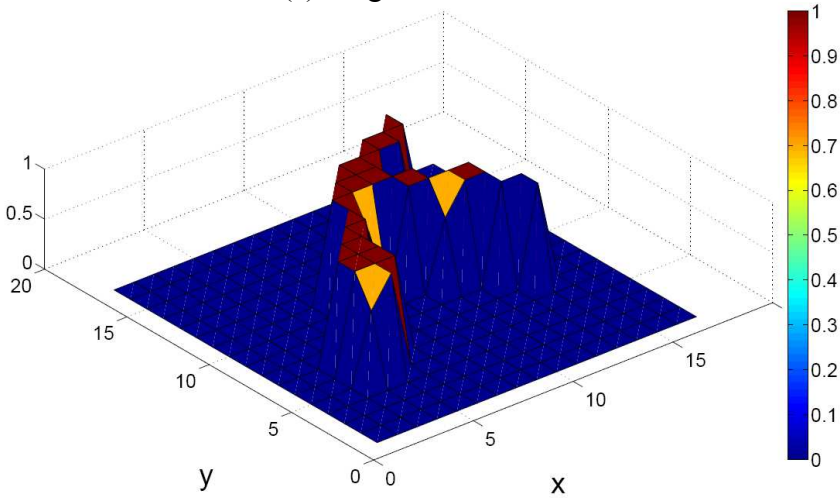


(c) Detected dense areas by mean-shift clustering (areas visualized in different colors).

Figure 6: Mean-shift clustering application example.



(a) Original 3D data.



(b) Final 3D Filtered data.

Figure 7: Examples of dense area segmentation after applying the filtering stage.

Inputs:
m: Original 3D data grid
r: Neighbourhood radius
Output:
filt_m: computed gradient value of each grid position

Algorithm:
aux ← *m*;
for each grid position (x, y) **in** *m* **do**
  $filt\_m_{x,y} \leftarrow 0$;
for each grid position (x, y) **in** *m* **do**
  if $m_{x,y} > 0$ **then**
  // compute gradient measure of grid points using *r*
    for (x', y') **in** $N(x, y, r)$ **do**
    // point (x, y) and neighbour points of it at distance ≤ *r*
      $aux_{x',y'} \leftarrow aux_{x',y'} - m_{x,y}$;

$$filt\_m_{x,y} = \frac{\sum_{(x',y') \in N(x,y,r)} \left| aux_{x',y'} \right|}{card(N(x,y,r))};$$

normalize values of *filt_m* in [0,1]

Figure 8: Pseudo-code for computing the gradient measure *filt_m* in the "shape filtering" sub-stage.

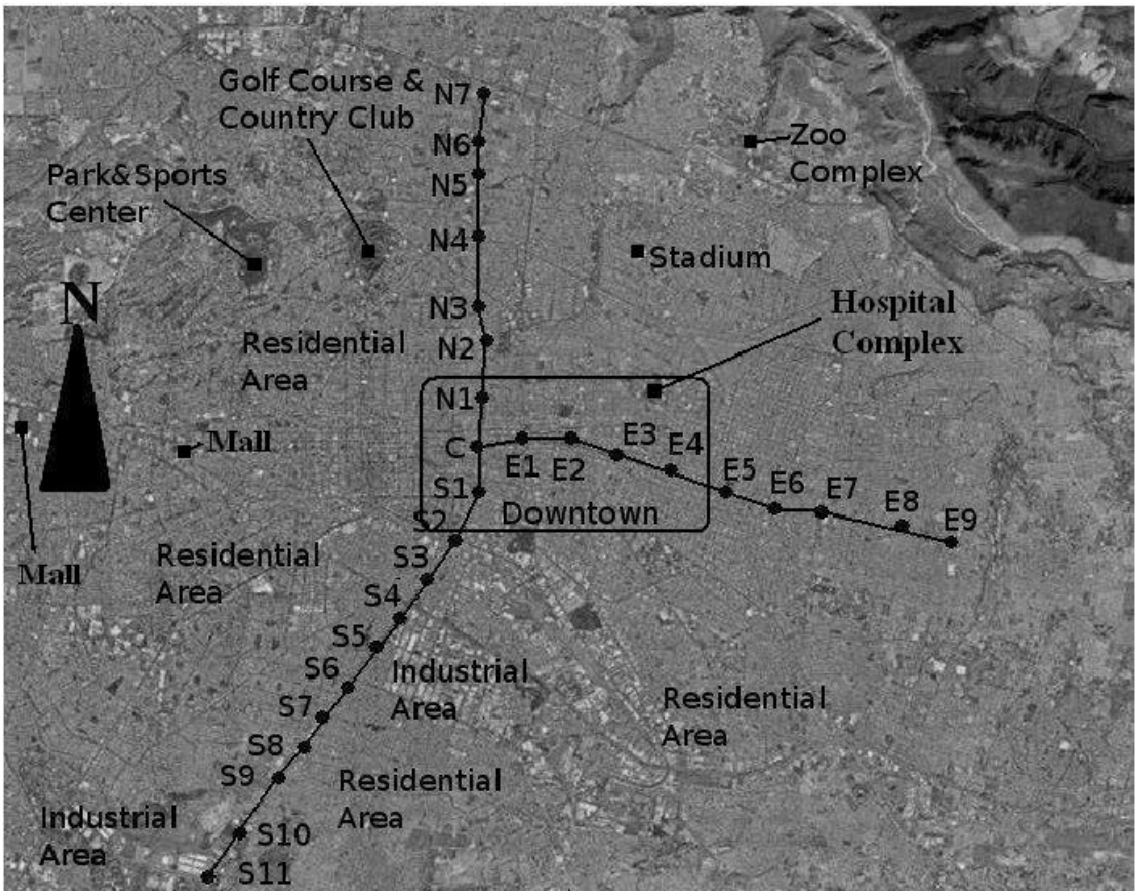
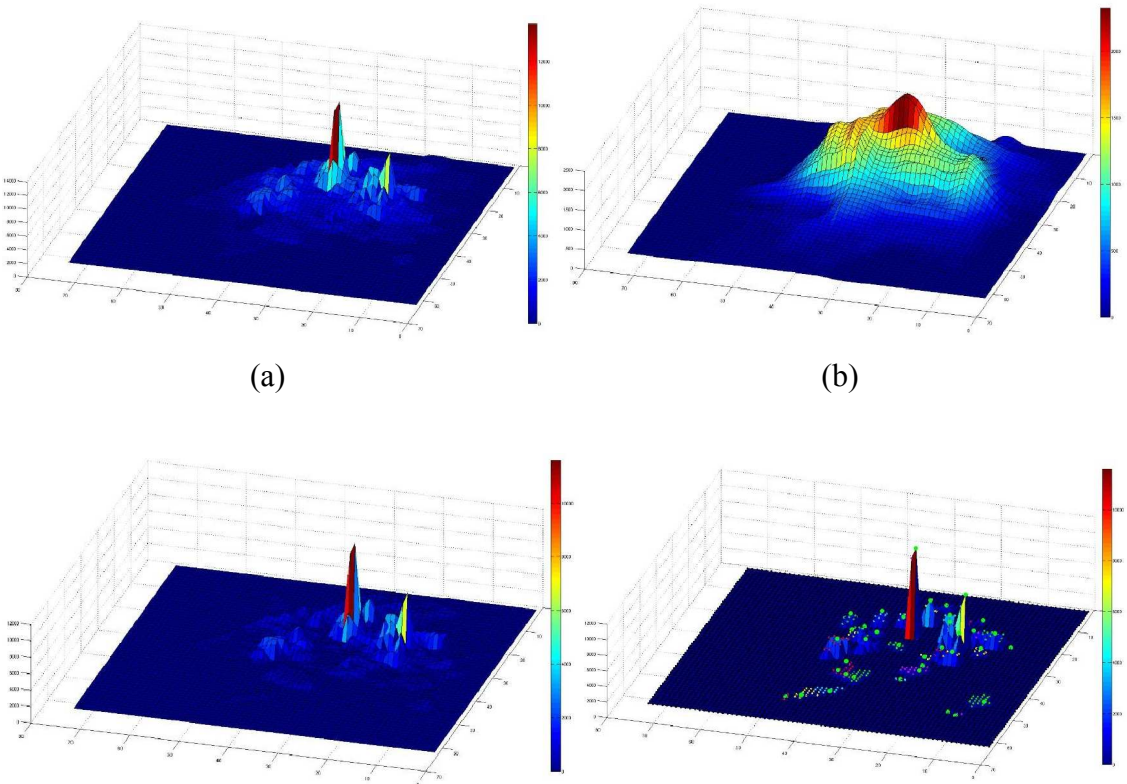


Figure 9: General description of the metropolitan area of Guadalajara.



(a) (b)

(c) (d)

Figure 10: Application of AdaptiveDAD to metropolitan area of Guadalajara: (a) data representation after constructing the grid; (b) baseline constructed with airPLS; (c) result of subtracting the baseline from original data and (d) representation of dense areas identified with their maximums after applying both mean-shift and gradient filter stages.

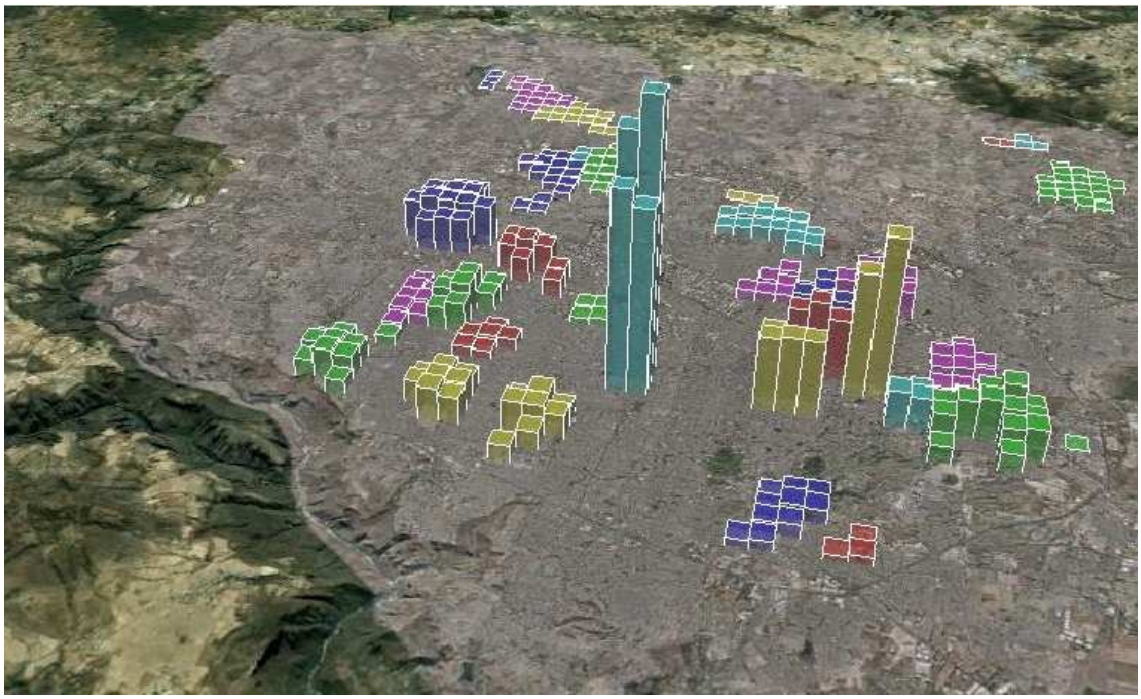
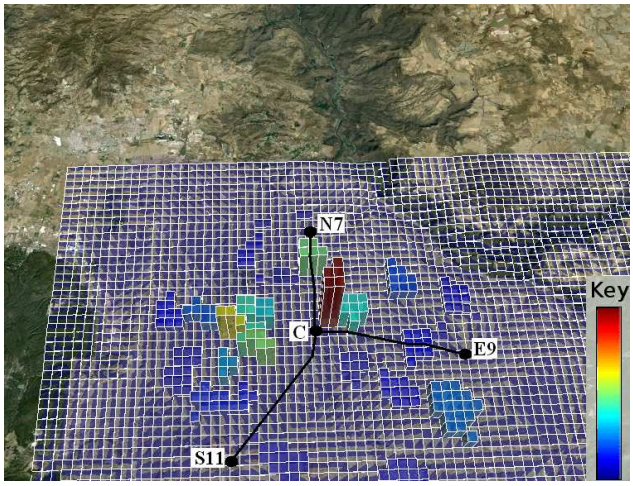


Figure 11: Representation of the results presented in Figure 10(d) over a map of Guadalajara.

(a)



(b)

Figure 12: Dense areas for the 6am to 9:59:59 time slot during weekdays identified with: (a) a traditional algorithm and (b) the AdaptiveDAD algorithm.

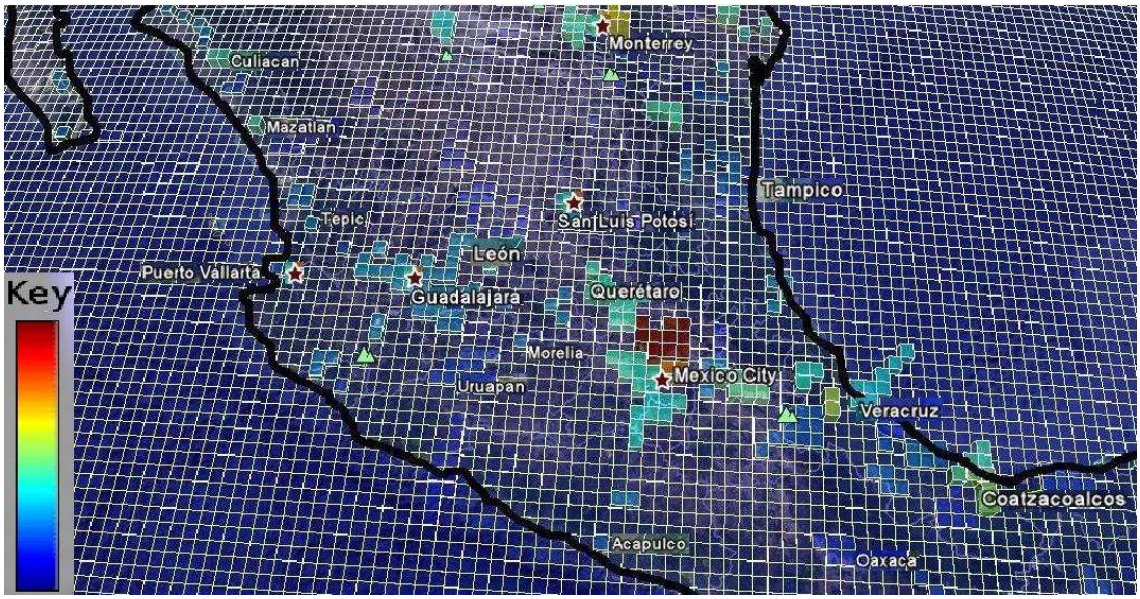


Figure 13: Representation of the dense areas identified at a national level with a grid of 0.2 degrees.

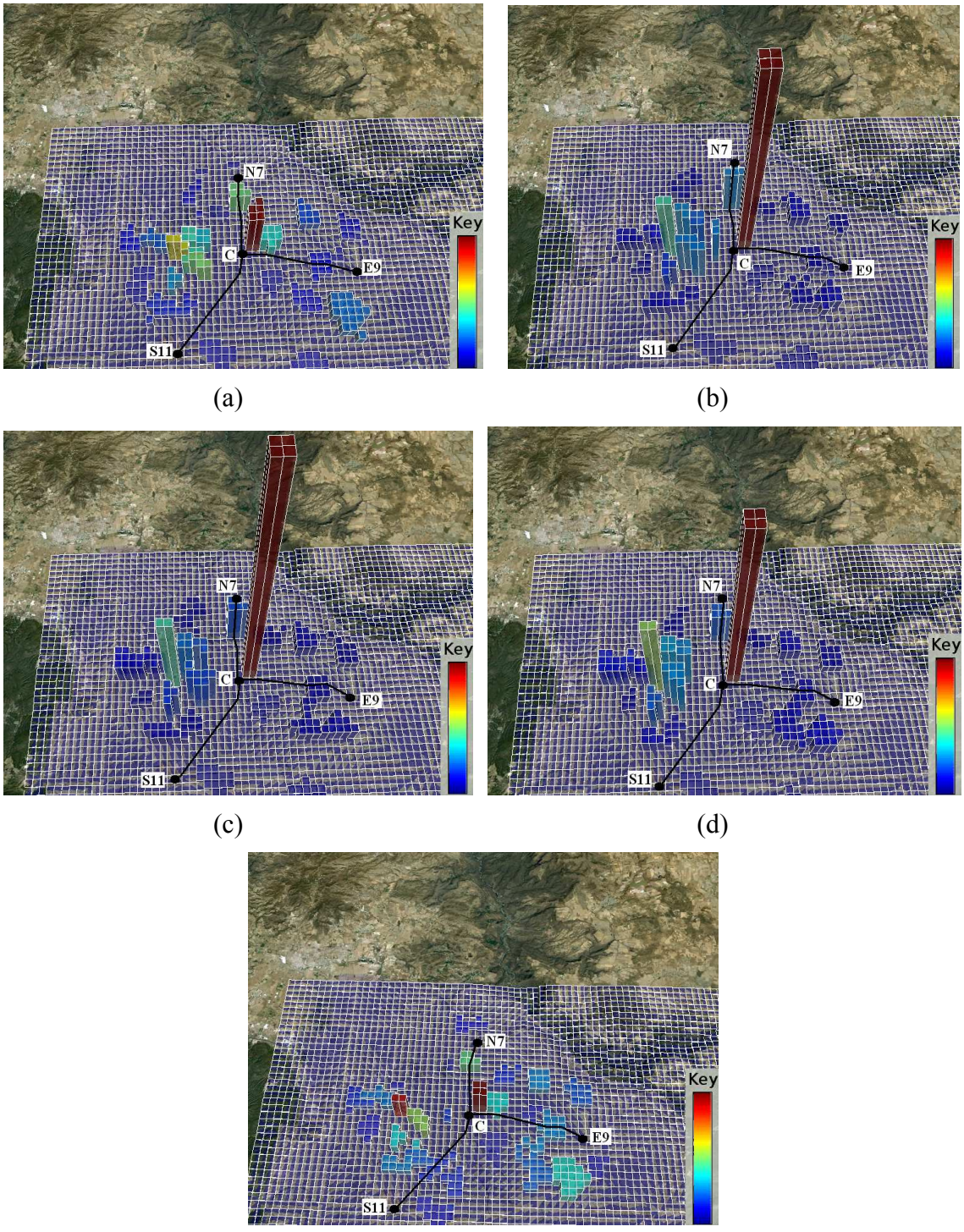


Figure 14: Evolution of dense areas identified by AdaptiveDAD in Guadalajara during weekdays for (a) 6am to 9:59:59am ; (b) 10am to 1:59:59pm; (c) 2pm to 5:59:59pm; (d) 6pm to 9:59:59pm and (e) 10pm to 1:59:59am.