**Statistics Netherlands**

Division Methodology and Quality
Sector Methodology Heerlen

*P.O.Box 4481*
*6401 CZ Heerlen*
*The Netherlands*

# Time patterns, geospatial clustering and mobility statistics based on mobile phone network data

**Edwin de Jonge, Merijn van Pelt and Marko Roos**

*Summary:* We explore a mobile phone call activity dataset for its possible use in official statistics. The dataset provides longitudinal, geospatial indicators that relate to economic and cultural activity. An analysis of regional clustering of call activity is conducted. We look at the mobility of mobile phone users by using logged call-events and compare the results of this analysis to official mobility statistics.

*Keywords: Statistics, Mobile phone data, Time patterns, Geospatial clustering, Mobility statistics.*

Remarks:
The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

*Project number:* *CBO-*
*BPA number:* *-CBO*
*Date:*

## 1. Introduction

The use of mobile phones nowadays is ubiquitous. People often carry phones with them and use their phones throughout the day. Instrumental for the infrastructure enabling the coverage for mobile phones, are mobile phone masts/towers, called 'sites' in the industry. Those sites are located at strategic points, covering as wide an area as possible. Strategic points mostly are high vantage points without any obstacles blocking the waves.

Much of the activity that is associated with handling the phone traffic, i.e. handling the localisation of mobile phones, optimizing the capacity of a site, handling billing information, is stored by the mobile phone company. So mobile phone companies record data that are very closely associated with behaviour of people; behaviour that is of interest to us as a statistical agency. Obvious examples are behaviour regarding tourism, mobility, commuting and transport. In spatial statistics for instance we have a clear understanding of the distribution of the night-time population, based on population registration data. The destinations and residences of people during day-time are topics of various surveys. Using data from mobile phone companies we should be able to provide additional and more detailed insight on the whereabouts and the activity of the day-time population

For our research we obtained a dataset from a mobile telecommunication provider containing records of all call-events (speech-calls and text messages) on their network in the Netherlands for a time period of two weeks. Each record contains information about the time and serving antenna of a call-event and an (scrambled version of the) identification number of the phone. In this paper we use the data description of Roos (2009). The contributions of this paper are the exploratory analyses of time, regional and mobility patterns.

Also a map containing information on the geo-locations of 20.000 sites was provided. We transformed this so called "cell plan" into an appropriate tessellation needed for geospatial analysis. We use absolute and relative call activity to detect deviating day patterns: We also experimented with different regional normalisations to detect regional clusters. In this paper we will use the region of "Eindhoven" as an example to describe the regional clusters. The data however is available for the complete country of the Netherlands.

## 2. Possible statistics based on phone data

Statistics Netherlands is responsible for collecting and processing data in order to publish statistics to be used in practice by policymakers and for scientific research.

Analyzing the mobile phone data is of interest for Statistics Netherlands because this data may prove a proxy for at least three important figures: mobility of people, population density and economic activity. These figures demand a different approach for this data, as we will show in this paper. The usage of cell phones is so wide spread that it may be possible to estimate population densities at a detailed regional and time resolution. Furthermore it is reasonable to assume that call activity in commercial or business areas is an indicator for economic activity.

We therefore see several opportunities for a statistical use of mobile phone data:

- Dynamic population density estimates: current population density is based on place of residence: where do people live? An equally important question for policy makers is which regions are densely populated during the day and year: how population density changes over time.
- Commuting: using (scrambled to protect privacy) cell phone ids and time and location data of a call event, it may be possible to estimate commuters mobility in predefined regions.
- Tourist statistics: tourist statistics are traditionally based on surveys. Many tourists visiting the Netherlands use cell phones during their stay in the Netherlands. The foreign cell phones enter the Dutch network as "roaming" and are registered as such. Using this "roaming" information it may be possible to estimate the number of tourists more accurately. Also figures about the duration of their stay in the Netherlands may be produced.
- Mobile phone data is very detailed in time and region compared to survey statistics. The data may be used for breaking down the traditional tables into more detailed statistics, e.g. daily and neighbourhood statistics. For instance the impact of special events could be distinguished.
- Early indicator for economic activity: the number of phone calls can be an indicator of economic activity in a certain region. Analysing these data might lead to an indicator that shows that economic activity for a certain region is changing.
- Classify regions as "residential", "commercial", or "business". The call activity of a region during the week for residential, commercial and business areas is different. It may be possible to derive a classification from the call activity profile of a region.

The main advantage of this mobile phone data is its detail in region and time: it offers opportunities to make detailed time series of detailed regions. Furthermore it may offer timely statistics: it is a matter of processing, transforming and analysing the data received from the provider.

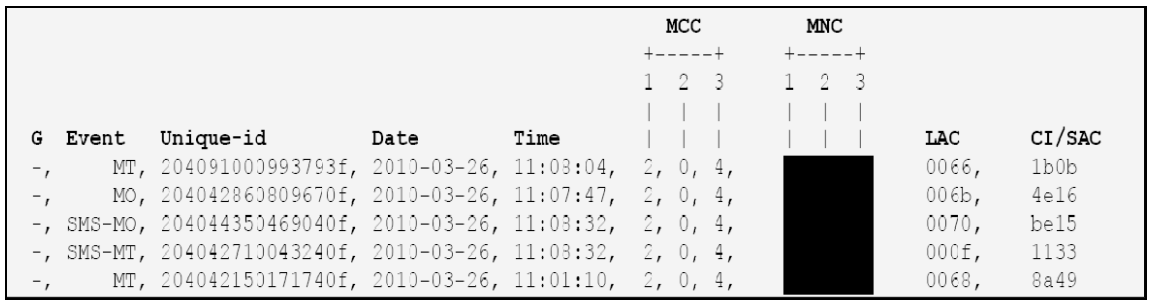
Before we can really use this data in a useful way, it should be established that call activity can be used to estimate the population density or economic activity accurately.

## 3. Data layout

### *3.1.1 Mobile phone data description*

We received all call- and text-messages events during a two week period from a leading phone network provider in the Netherlands. For each call the starting time, the site location and a scrambled version of the unique phone id (IMSI) are available. The data we received is as described in Figure 1.

With some adjustments in its systems the mobile phone company was able to channel the selection of the data on a hard disk, which was subsequently encrypted. We then imported and decrypted the data onto a secured network.



| G | Event | Unique-id | Date | Time | MCC | MNC | LAC | CI/SAC |
|---|---|---|---|---|---|---|---|---|
| -, | MT, | 204091003993793f, | 2010-03-26, | 11:03:04, | 2, 0, 4, | [redacted] | 0066, | 1b0b |
| -, | MO, | 204042863809670f, | 2010-03-26, | 11:07:47, | 2, 0, 4, | | 006b, | 4e16 |
| -, | SMS-MO, | 204044350469040f, | 2010-03-26, | 11:03:32, | 2, 0, 4, | | 0070, | be15 |
| -, | SMS-MT, | 204042710043240f, | 2010-03-26, | 11:03:32, | 2, 0, 4, | | 000f, | 1133 |
| -, | MT, | 204042150171740f, | 2010-03-26, | 11:01:10, | 2, 0, 4, | | 0068, | 8a49 |

**Figure 1: Layout of mobile phone data**

Figure 1 provides an example of the transactional data we received from the mobile phone company. The MNC (Mobile Network Code) column is blacked out because it reveals the mobile phone operator.

With the data we can relate an event (sending/ receiving call or text message) to a unique phone, a date and time, and to a site geo-location. It is important to note that each unique ID remains stable throughout the period contained in the dataset. This means that if a certain IMSI makes a call at a certain site 'x' and later during the day a call at site 'y', we are able to recognize this in the dataset.

The data were recorded during the period of April 29th 2010 until May 9th 2010. In this period, two significant events took place: Queensday on April 30th and Liberation Day on May 5th. School holidays were from April 30th until May 9th.

Event data for all cells in the Netherlands for this period were recorded and made available.

### *3.1.2 Spatial distribution of sites and their cells*

A very important aspect of the mobile phone data is the spatial distribution of the calls that were made. The call data contains the "cell id" of the cell that made the connection to the network, so we need to know the locations of each cell. This is called a cell plan. A cell plan is dynamic, it changes often because mobile providers add new antennas but also have mobile antennas that are used to cope with high demand. In our experiment we assume the cell plan is static and we ignore cells that are not in the cell plan. We also simplified the cell plan by grouping all cells of the same site. However, we differentiate between different technologies (UMTS/GSM) because the sites partly overlap.

Sites are coded in the data with cell id (CI/SAC). Mobile phones connect to the network via an intricate network of cells. Cells receive and beam phone signals serving a specific area, basically resembling a slice of a pie. Cells are combined on a site; usually in groups of three (see Figure 2).

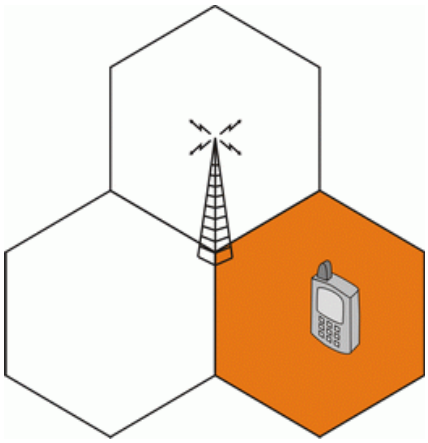


**Figure 2: Cell and sites**

One of the goals of our study was to plot geospatial activity of mobile phone usage on a map. From the mobile phone company we received a number of files, each describing the cell locations based on different formats, such as WGS84 and the Dutch geographical projection 'RijksDriehoekstelsel' (RDS) and postal code. The 3G cell plan (see paragraph 3.1.3) is provided with polygon shaped cells, but these are an approximation of the serving area of a cell: the coverage of a cell depends on weather conditions. Furthermore cells can offload calls to neighbouring cells when their maximum capacity is reached.

We opted for RDS because publications based on postal codes pose copyright problems and still have to be translated into RDS coordinates. Furthermore, postal areas and cell plan do not coincide nicely. With RDS we could also make a direct translation to geospatial neighbourhood information published by Statistics Netherlands.

A map based on the cell plan for each different technology (see paragraph 3.1.3) was created.

### *3.1.3 Infrastructure technologies*

Mobile phone companies operate three different infrastructure technologies. Those technologies are GSM, UMTS (or 3G) and Dual. The Dual (or Dual band) technology is negligible, because it is outdated and hardly used. GSM handles 'normal' telephone activity (calling and receiving speech-calls and text messages). UMTS handles both normal telephone activity as well as high volume data traffic (video calling, mobile internet). Each technology uses different cells, with different ranges and angles. UMTS cells are often, but not always, located at the same site as GSM cells. Table 1 provides an overview of the number of cells per technology. Some sites locations contain multiple technologies and most sites have 3 or 6 cells attached. The number of site locations is not equal to the sum of the sites in the table because of multiple technologies on one site location.

| **Technology** | **Sites** | **Cells** |
|---|---|---|
| GSM | 3278 | 9157 |
| UMTS | 3070 | 9245 |
| Dual | 396 | 725 |

**Table 1: Sites and cells per technology.**

### *3.1.4 Privacy issues*

Our mobile phone data relate directly to behaviour of individuals. The data are anonymized, meaning that the IMSI's are scrambled and no direct relation to a specific person can be made.

This was a specific demand of the mobile phone company, but even if it had not been their demand, we would have made it one of the requirements. As a statistical office we rely heavily on the trust of people that we handle data while respecting the privacy of people.

Despite our efforts to make the data anonymous, we are dealing with sensitive data. Our current data set is privacy safe, but it is wise to regard the ethics of future research.

## 3.2 Patterns in the overall data

In our research we want to explore the data provided by the telecom provider. We conjecture that the data could help finding answers to the following research questions:

Hypotheses:

- Economic activity is related to regional call activity.
- Mobility of persons is measurable using call traces containing location data
- Estimation of number of persons per time and location is possible with number of calls (per time and location) (or call activity can be used as a "proxy" for number of persons)
- Calling behaviour of people can be classified to number of calls in a time frame / frequency of calls.

Whether the data contains suggestions for these questions is subject of research in this article. Not all questions following from the research questions will be answered in full. Where we do not get a full answer we will give reasons for not being able to find one. Let us first take a look at the data.

The data provided were 14 days in a sequence: April 26$^{th}$ untill May 9$^{th}$ 2010. This time interval is special because it contained two national holidays. This was good in a sense that the dataset contains, week days, weekend days and holidays, but the down-side is that the time interval is too short to derive a realistic week day pattern. It was, nevertheless, an interesting experimental data set to explore several research questions.

The overall call intensity in the two week period is shown in Figure 3.

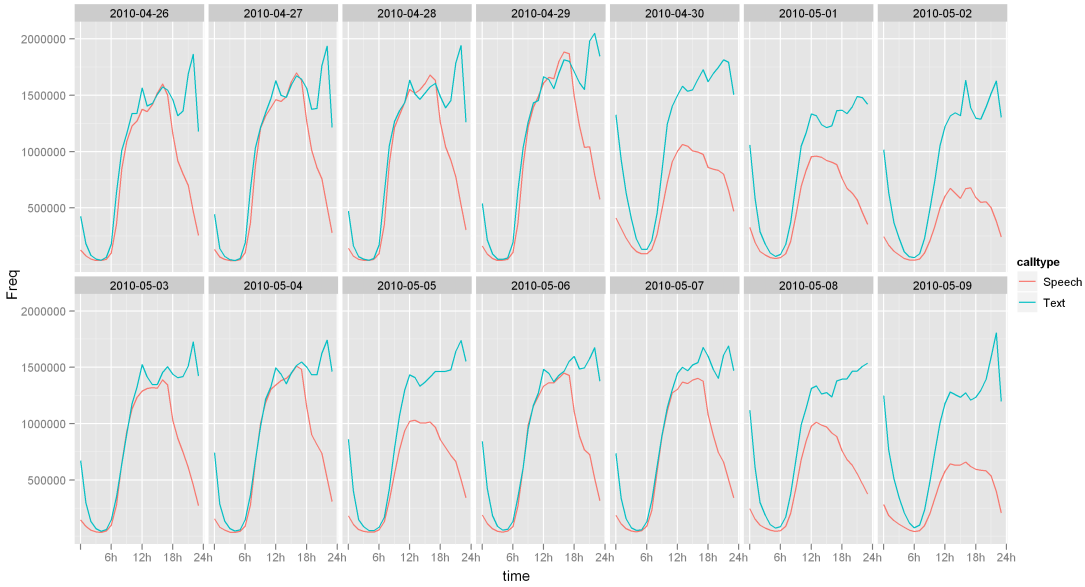



**Figure 3: Speech and Text call frequency from April 26th till May 9th**

Clearly visible in the (speech) activity are the working days, weekend days (May 1st, 2nd, 8th, 9th) and holidays (April 30th and May 5th). This is a clear indication that call activity is related to economic activity. Figure 3 also shows the differences in activity between speech calls and text messages (SMS); text messages have several peeks and are sent even late at night.

The phone data show a stronger working day pattern than the text data. Ás a result, working day patterns can be identified better when text messages are removed from the data. For this reason, we decided to use the speech data only when analyzing working day patterns.

Exploring the data, we see that text message data look less sensitive for working day, weekend day and holiday patterns. However text messaging in the night before a weekend day or holiday stops later, which is an indication that text messaging is related to leisure time. Remarkable are the three activity peaks around 12:00h, 18:00h and 23:00h for each day. The last data peak may be useful when classifying regions as residential or not-residential: it is reasonable to assume that a (daily) peak at 23:00h occurs when the mobile phone owners are at home.

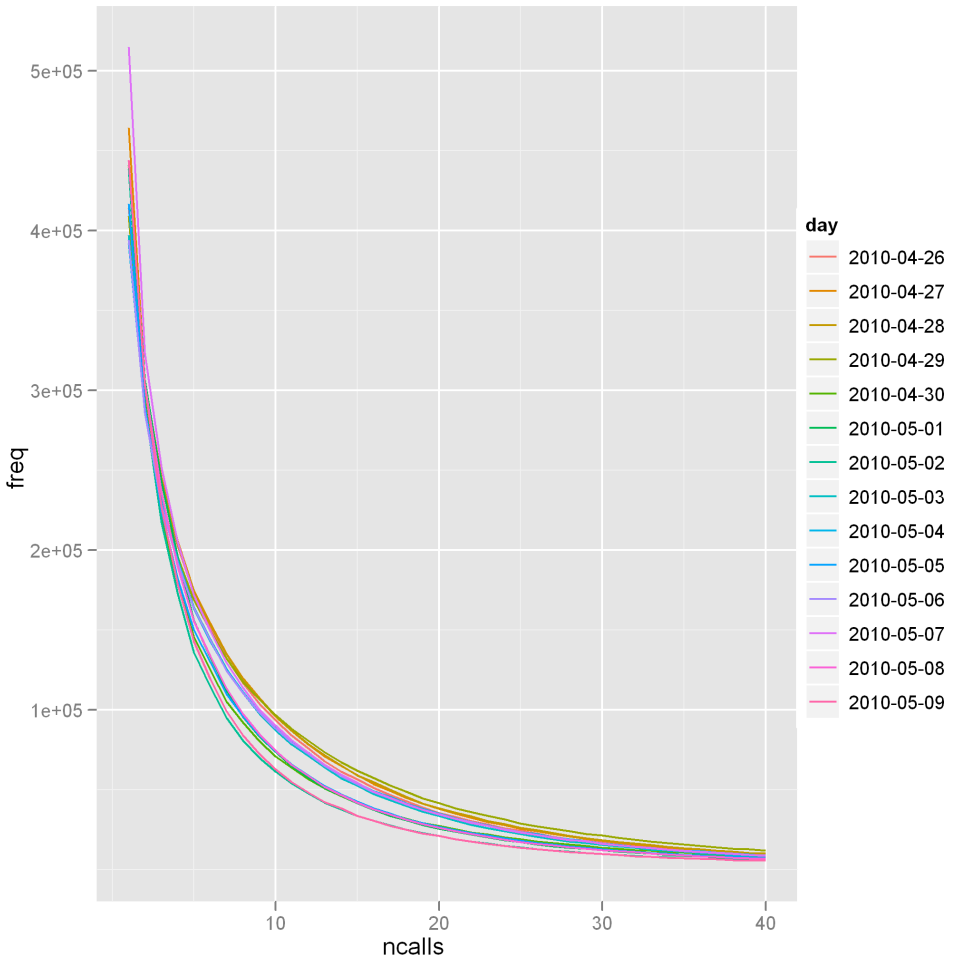



**Figure 4: Frequency of calls per phone per day**

Figure 4 shows the distribution of the number of calls per phone per day. In the figure the number of calls (ncalls) is restricted to 40, which covers more than 99% of all calls. Remarkable is the smoothness of the frequency distributions and their similarity. It may be possible to use this distribution to estimate the phone density based on the number of calls per site.

## 4. From data to statistics

The raw mobile phone data need to be processed for making statistics about population density and economic activity. The raw phone data contain calls per site location per unique (scrambled) phone id per datetime in seconds. Both time and location need to be transformed before statistics can be made.

The time resolution in seconds is too precise to be used for statistical analysis, so we aggregate the time variable in a histogram for all calls. We experimented with five minutes and 60 minutes time bins. For the clustering, we settled for a sixty minutes time bin.

Special care is needed when estimating population density: in stead of number of calls, the unique number of calls has to be counted otherwise frequent callers will bias the estimation. The removal of double entries per time interval is laborious, since there are approximately 5

million different phones in our dataset. We refer to this activity in this paper as "undoubling" phone ids.

We are interested in the spatial distribution of number of calls-events, but the location data contains the fixed positions[1] of sites and not of the mobile phones. These positions cannot be used directly to plot the spatial distribution of the data because towers have a fixed location and the location data cannot be treated as generated by a random point process. Each site serves an area with a different size and location. We define "call-event density" $d_i$ as the distribution of activity $a_i$ of site *i* over its neighbourhood $v_i$ as

$$d_i = \frac{a_i}{area(v_i)}$$

To determine neighbourhoods $v_i$ we make the (simplified) assumption that each site serves a unique region and a phone connects to its nearest site. Based on these assumptions a Voronoi diagram (*Voronoi 1907, Dirichlet* 1850) can be made that defines the neighbourhoods of the sites. Figure 5 shows the Voronoi tessellation of the Eindhoven area, where dots are the GSM sites and the thin grey lines are the Voronoi area borders. The white lines represent municipality borders in the Eindhoven area. Notice that sites are not uniformly spaced, as the area becomes more rural, the density of dots also becomes lower.

An important issue with the mobile phone data for it use in population estimation is censoring: the dataset contains only calls made on mobile phones and not the location of active mobile phones. This is extra important for mobility statistics which we will discuss in paragraph 6: many changes in location of mobile phones are not registered, only those who make subsequent calls.

[1] There are mobile sites, but even these cannot be treated as random.

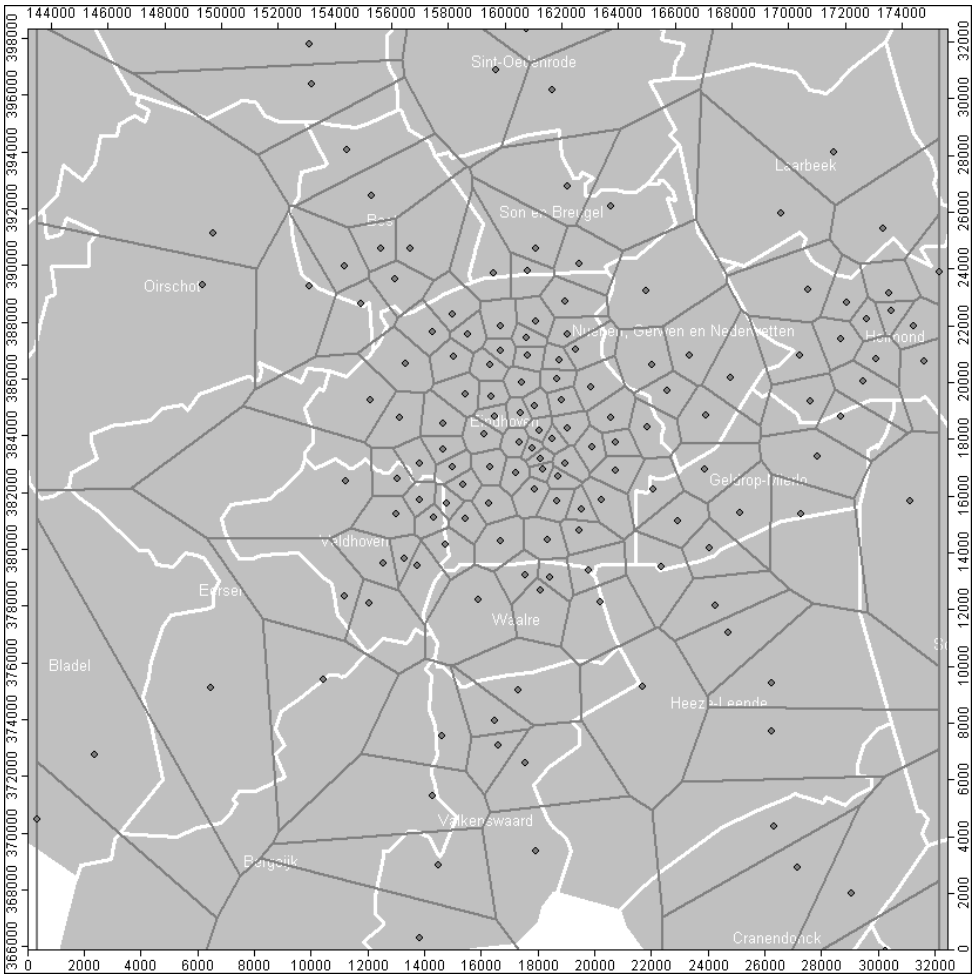



**Figure 5: Voronoi tessellation of GSM sites in Eindhoven area**

A mobile provider offers multiple mobile technologies, GSM, UMTS, so there are separate networks for each technology. The location of their sites is often different. The networks and their call activity have to be treated separately. Modern phones can connect to all types of networks, so the mobile phones on the different networks are not separated populations. In the analysis of the spatial distribution of the call activity, the activity of all networks has to be combined.

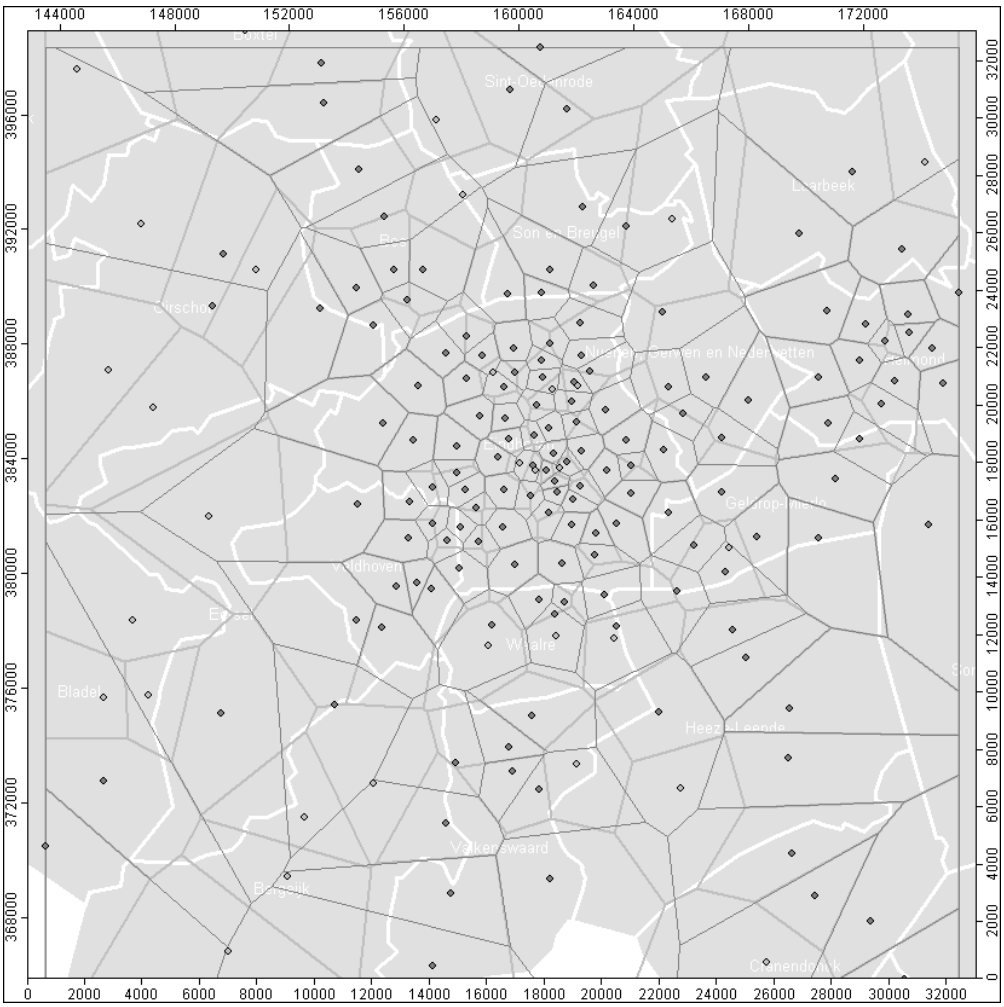
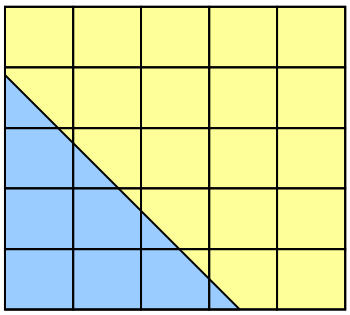



**Figure 6: Voronoi tessellations for UMTS and GSM technologies**

### *4.1.1 Grid: combining areas*

Connecting socio-economical data from official neighbourhoods with mobile phone activity would allow for spatial regression GSM and UMTS have different Voronoi tessellations, which are also different from official geographical tessellations such as municipalities and neighbourhoods. These tessellations can be combined by "gridding" the data . We place a grid over all tessellations and assign each square of the grid the according value of the tessellation.



Instead of aggregating on site areas (which are often different in shape and size) aggregation is done on grid points.

To distinguish areas with a high density of activity from areas with a low density we plotted the log density with a spectrum scale (Rheingans, 2000). For maps with a different theme (e.g. the relative calling density) we choose different starting and ending colours.

**Figure 7: Call activity during the day**

To compare the deviation of activity within a site area against some kind of mean (e.g. the mean of all time frames within a day, the mean of a specific time frame over working days, the mean of a specific time frame over sites), we used a double colour spectrum scale.

## 5. Clusters in the data

### 5.1 Data patterns

One desired result of the data set is to categorize areas to 'residential and not residential' and 'commercial and not commercial' based on the call activity of the phone data. The hypothesis is that residential areas have a different call activity pattern then commercial or business areas. It is to be expected that during normal working hours commercial and business areas are more active then residential areas and, therefore, also the speech call activity. Alternatively, it is reasonable to assume that during typical leisure hours (holidays and Sundays) residential areas are more active then business areas.

#### *5.1.1 Day patterns*

If working day patterns in mobile phone data exist, they should manifest themselves in the time-series for the number of calls during week days, weekends and holidays.

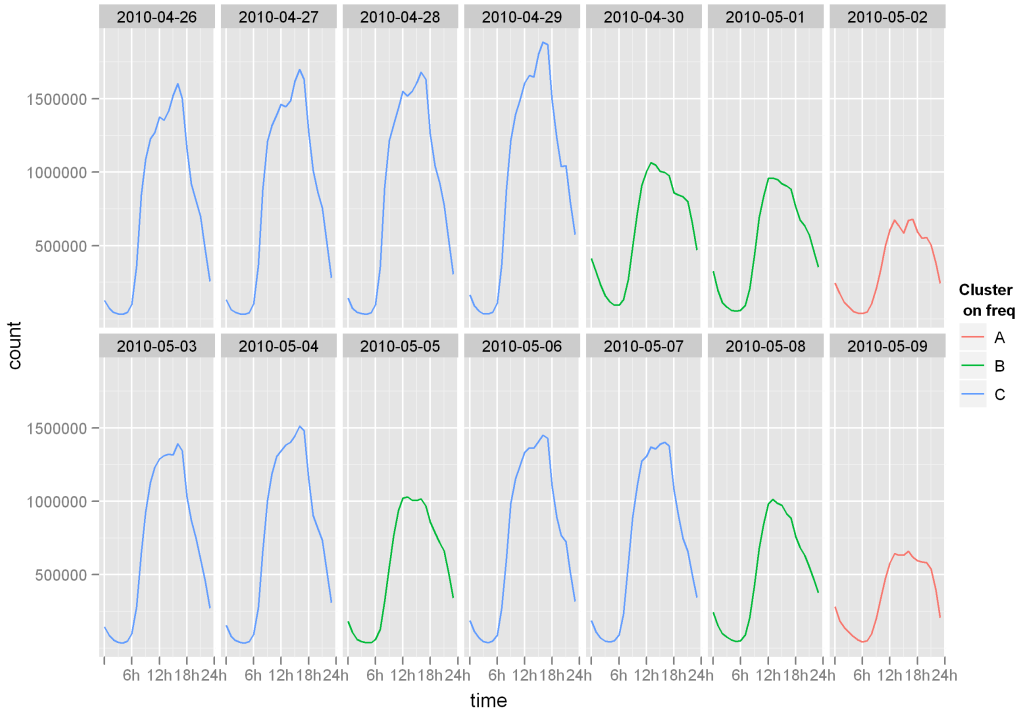



**Figure 8: Day cluster of Speech call frequency**

Figure 8 shows the time series of the number of calls, which contains clear differences in day patterns. To determine day pattern clusters we applied a kmeans clustering (MacQueen, J. B. (1967) which resulted in three distinct day clusters: Week days, Saturdays/holidays, Sundays. This strongly suggests that mobile speech call activity is related to economic activity. In Figure 9 a clustering on the relative frequency per day is shown. This clustering is nearly identical to the clustering on the absolute one. The difference is that April 30$^{th}$ is clustered together with Sundays.

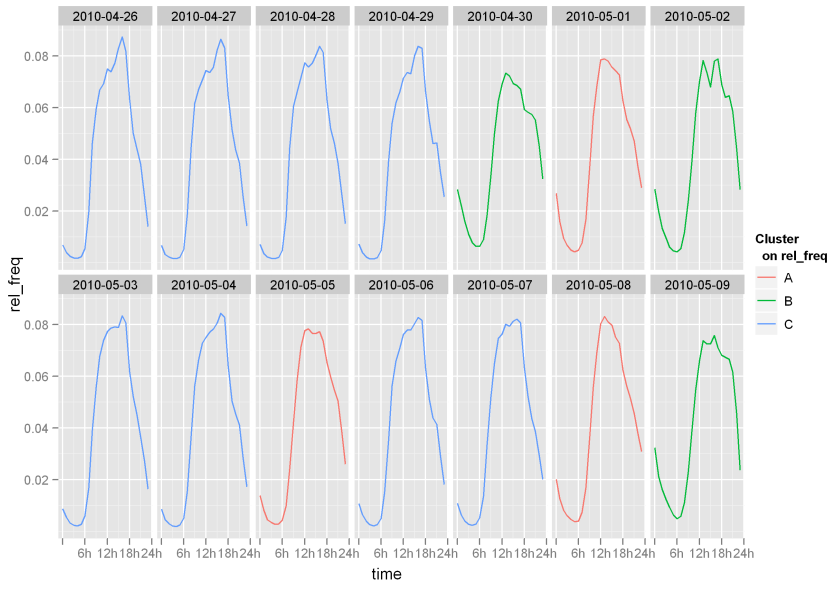



**Figure 9: Day cluster of relative Speech call frequency**

The accuracy of the resulting clustering is relatively low; the data set is restricted to two weeks. Since these two weeks accidentally contain two holidays, they are different from

average weeks. For a good weekday pattern a larger time span and a bigger data set is necessary.

### *5.1.2 Regional patterns*

To detect clusters of regions that are residential, commercial or business areas, a call activity pattern or "call signature" must be estimated for each region and these signatures must be clustered. We did some experiments with clustering regions and checking if they are commercial or residential, but for further conclusions a more detailed cell plan is needed: current Voronoi cells are too large to align with official neighbourhood regions.

## 6. Mobility

Statistics Netherlands produces the statistics 'Traffic and Transportation' and 'Survey Mobility in the Netherlands' in which the mobility of commercial transporters and Dutch inhabitants is measured. These statistics currently rely on surveys for collecting their data. Mobile phone data is an interesting additional source of information, because we can calculate travel distances of mobile phone users, which may be a good proxy for the mobility statistics. Furthermore mobile phone data could reveal high mobility geographical areas, which is of interest to policy makers. The advantage of using phone data is that no survey is needed, the data does not suffer from human memory effects (call events are logged on automated systems) and there are many observations. However the representativeness of the data and the censoring of location present methodological problems for analysis.

We compare mobile phone travelling distance with the official statistic and discuss the various methodological issues.

The mobile phone data contains the location and (scrambled) phone id of each call. Calls with the same phone id can be used to estimate a lower boundary for the travelling distance. For each mobile phone *y*, we define a vector ( $sv_{x,y}$ ) containing in chronological order all measured site identification codes $s_i \in S$ during day x. $sv_{x,y}$ contains points on the travelling route of mobile phone user *y* on day *x*. We estimated a lower boundary (see appendix I for the explanation) of the travelling distance using straight line estimations. Calculating site-vectors for all 5 million mobile phones was too computationally intensive, so we used two different samples and calculated the travelling distance for one fixed day *d*.

Sample I is sampled from the entire population of phones (IMSI's) active on day *d*. Most phones call once a day, for these phones we cannot calculate a travel distance.

Sample II is sampled from the active mobile phone population with multiple calls on day d, from regions around important Dutch cities. It contains "high mobility phone users". Both samples contain 2000 IMSI's and are summarized below.

| Sample I: distance (km) 2010-04-27 | |
|---|---|
| Min | 0.0 |
| 1st quartile | 0.0 |
| median | 5.6 |
| mean | 22.8 |
| 3rd quartile | 25.0 |
| max | 484.9 |

| Sample II: distance (km) 2010-04-27 | |
|---|---|
| Min | 3.9 |
| 1st quartile | 43.7 |
| median | 78.7 |
| Mean | 99.6 |
| 3rd quartile | 133.5 |
| max | 571.2 |

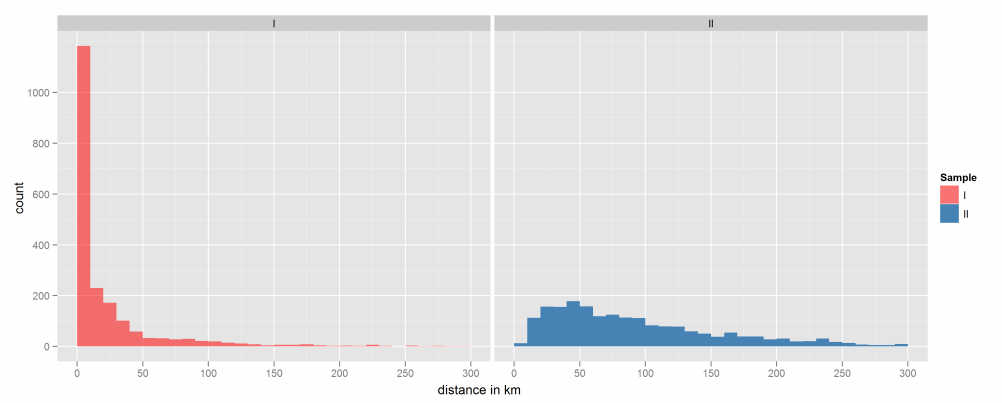



**Figure 10: Histograms of the travelling distances in sample I & II**

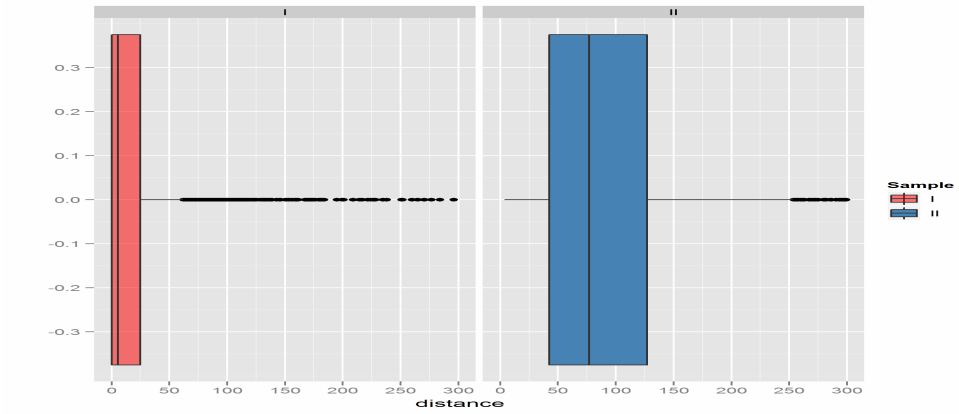



**Figure 11: Boxplots of the travelling distances in sample I & II**

Figure 14 shows that the travelling distance in both samples is (very) skewed, suggesting that the arithmetic mean travelling distance is an inaccurate measure for describing average travelling distance.

The mean of sample I is 23 km, which is, as expected, an underestimation compared to the official average travelling distance of 32 km per person per day (OViN (2011)).

The spatial distribution of the travels is interesting. A map of the Netherlands was plotted with all travels made by mobile phone users in each sample. For the 'highly mobile' mobile phone users (type II) this results in an insightful map (Figure 12) revealing the major cities, highways and train connections.

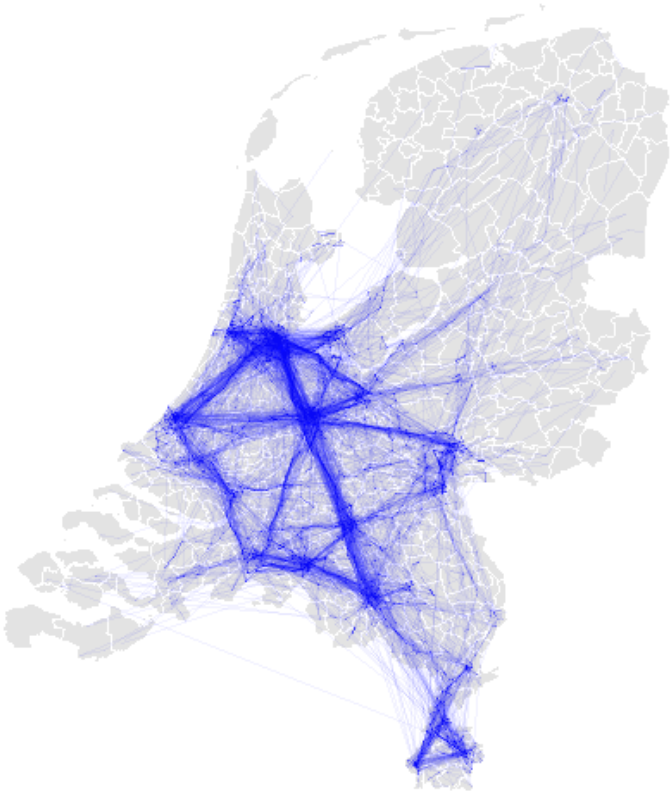


**Figure 12: Mobility of 'highly mobile' phone users (sample II)**

The map suffers from the sampling method we used for sample II: the north of the Netherlands is empty because the regions in the north are underrepresented in this sample.

In the South we see an interesting connection that suggests the existence of 'Bandstad' (Maastricht, Sittard, Heerlen, Luik/Liege and Aachen).

Figure 13 shows for sample type I that phone user travel across the entire country, although there is no clear pattern appearing.

**Figure 13: Mobility of mobile phone users (sample I)**

The estimation suffers from censoring: only locations where a call is made are registered and we use only phones with multiple calls per day. However there is a counteracting effect that people tend to call before, during and just after a travel: travellers may phone more frequently. The estimation also suffers from the Euclidean distance approach: a routing planner will give higher travel distances (for a possible method to improve the estimation see appendix II).

A part of censoring could be reduced by assigning each phone a 'home-location', so all phone calls can be used in the calculations of travel distance. A home location would also allow a regional break down of mobility statistics.

## 7. Implementation/processing phone data

The total size of unpacked data was about 67 Gigabyte. The data were processed using R. For our purposes we needed to create tables for each site and day with the total activity during a five minute frame. Processing the data from raw CSV file format into usable tables was one of the most time consuming aspects of the project.

The number of calls/text messages was 40 million a day, the total data around 600 million records. Initial scripts to transform raw data into R-format files and add site information took several days per MSC to run. Several times we ran into memory and storage problems on the infrastructure (virtualized desktop and network storage) we used.

We tried to enhance the performance by using the FF package of R, but this required a significant amount of tuning and a lagging performance in a virtualized desktop and network

area storage. Furthermore, because the processing code had to be 'chunked' because data did not fit into memory. Tests on a stand alone machine and local hard disk turned out to be several times faster.

## 8. Recommendations for further work

This study explored a mobile phone data set for several uses for official statistics. The findings confirm that mobile phone data may be of use to statistical topics varying from economic activity, tourism, population density to mobility and road use.

We did not explore all possible uses of the data: we would like to explore its use for tourist statistics, traffic and "connectedness". Mobile phone data may prove to be even richer in their application. In a recent presentation (Ahas (2011)) pointed out the many detailed possibilities of roaming information on tourism activity. Social cohesion, a statistical measure based on the Lisbon Strategy, could be made more insightful with data on the geographical distribution of calls; who's calling who (more specifically: from where to where?).

For further investigation a more detailed cell plan model is needed increasing the geographical resolution with a factor 3. Furthermore to determine stable and reliable economic patterns a dataset spanning several months is needed. Linkage of the data with the Dutch population register would give more insight in background variables of mobile phone users.

A major limitation of the data is that we currently do not know how the mobile phone dataset represents the (total) population of the Netherlands. Another limitation is censoring of locations: only a subset of the excursions is identified.

If we want to derive population densities from call activity we have to know the correlation between calling activity and population density. If a certain number of people call during a given moment of the day, how do we know what percentage of the total number of people present in that area actually made or received a call? One way is to use the mobility statistic to ask questions concerning phone use. These data can be used to assign basic calling patterns to areas and time frames (morning, afternoon, evening). Another possibility is to fit calling activity per region with the official residential population density.

Another idea worth investigating is to use the data to break down aggregated estimates of existing statistics in time and region. Based on data from a survey with a known sampling design and a rich set of linked background variables, the mobile phone data could be used to improve resolution in time and place. The table figures of the statistic would in this case be considered accurate and can be used as margins. The more detailed mobile phone data are then fitted towards those margins.

**Reference**:

Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H., Mark, Ü. (2007). "Mobile positioning in space-time behaviour studies: Social Positioning Method experiments in Estonia". Cartography and Geographic Information Science **34(4)**: 259-273.

Ahas, R et al. (2011), "Mobile telephones and mobile positioning data as source for statistics: Estonian experiences".". presentation for NTTS 2011 Brussels.

De Bruijne, A., Van Buren, J., Kösters, A., Van der Marel, H. (2005). "*De geodetische referentiestelsels van Nederland, Geodetic reference frames in the Netherlands"*, http://www.ncg.knaw.nl/Publicaties/Groen/pdf/43Referentie.pdf

Dirichlet, G. L. (1850). "*Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen"*. Journal für die Reine und Angewandte Mathematik, **40**:209-227

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". Numerische Mathematik **1**: 269–271

Lisbon Strategy (2011): http://en.wikipedia.org/wiki/Lisbon_Strategy

MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". 1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297

Newman, M. E. J. (2005). "Power laws, Pareto distributions and Zipf's law". Contemporary Physics 46 (5): 323-351.

OViN (2011), *Mobiliteit per regio naar motief en algemene kenmerken.* Statline table (Statistics Netherland). http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=37727&D1=1&D2=0&D3=0&D4=0&D5=0&D6=10-23&VW=T

R Development Core Team (2009). R: A language and environment for

statistical computing. R Foundation for Statistical Computing,

Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Voronoi, G. F. (1907). "*Nouvelles applications des paramètres continus à la théorie des formes quadratiques »*. Journal für die Reine und Angewandte Mathematik, **133**:97-178, 190

List of abbreviations:

3G: Third generation mobile telecommunications

CI: Cell Identifier

GIS: Geographic Information System

GPRS: General Packet Radio Service

GSM: Global System for Mobile Communications

ID: Identifier

IMSI: International Mobile Subscriber Identity

LAC: Location Area Code

MCC: Mobile Country Code

MNC: Mobile Network Code

MO: Mobile Originated

MT: Mobile Terminated

RDS: RijksDriehoekstelsel

SAC: Serving Area Code

SMS: Short Message Service

UMTS: Universal Mobile Telecommunications System

WGS84: World Geodetic System 84

Appendix I

Notice that the site vector underestimates precise route of *y:* it contains only locations of call events so it is unknown what other locations were on the route of *y*. We illustrate this with a graphical example. In Figure 14 the real route has colour green, the estimated route is of blue colour. It is clear that the blue route is an underestimation of the real travelled distance. There are at least three reasons that contribute to the underestimation. First, a straight line between the points $s_1, s_2, s_3, s_4, s_5$ is not an accurate representation of the real route: it takes the shortest distance possible while the real travelling distance using transportation in general is larger. Second, site vectors of length one are currently ignored. However if we knew the "home region" of a mobile phone we could derive extra travelling distance. Third, the set of location points is limited to call-events.

.

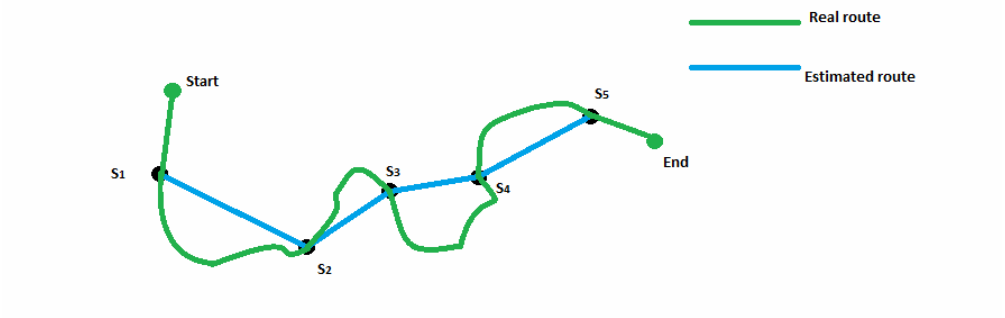



**Figure 14: Difference between real and approximated route**

Appendix II

We thought of a methode for improving the determination of the length of the travelled route of a mobile phone user. In this mehtode we construct a graph using the cellplan. The nodes of the graph are the locations of the sites in the tessellation. There is an edge between two nodes in the graph when the Voronoi tessellation areas of the sites represented by the nodes are adjacent. The labels on the edges represent the straight line distance between the neighbouring nodes. When travelling, one always has to go from one area to a neighbouring area. So to completely record a route travelled one needs to know all the edges of the graph used in the route. Suppose the travelled route is like the route in Figure 14. This route has site-vector $s_1, s_2, s_3, s_4, s_5$. Let's zoom in on the first part of the route between $s_1$ and $s_2$. We will look at the Voronoi tessellation surrounding these points and construct a graph useable for determining the distance.

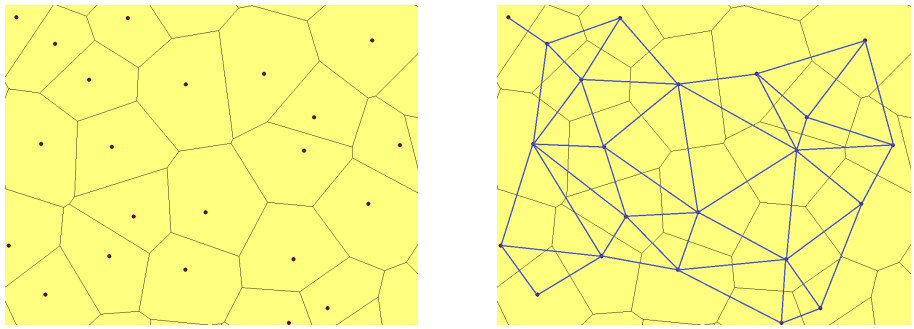


**Figure 15: Left: Voronoi tessellation of cell plan,**
**Right: graph constructed on tessellation**

In Figure 15 (left) one can see a Voronoi tessellation. The dots are the site locations and the lines are the borders of the site serving areas based on the Voronoi tessellation. In Figure 15 (right) the graph useable for determining the distance is shown. The site locations (dots) are connected by the edges (blue lines) of adjacent areas surrounding the dots. Suppose the upper most left point is $s_1$ and the lower most right point is $s_2$. One can use this graph in combination with the shortest path algorithm (Dijkstra's algorithm) (Dijkstra (1959)) to estimate the distance of the route traversed between $s_1$ and $s_2$. When the Voronoi tessellation areas become smaller this method becomes more accurate.

The problem of underestimating the traversed route can be made less severe by using extra information. When a route becomes larger the deviations of the estimated route from the real route can become larger. This is especially true for routes with a small number of internal points.

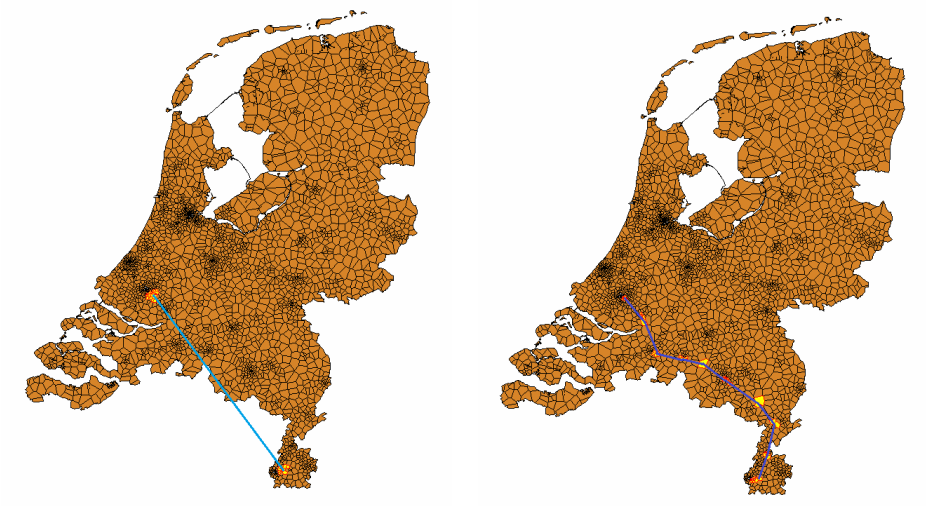


**Figure 16: Left: direct route (2 internal points)**
**Right: more realistic route (more internal points)**

In Figure 16 (left) a route is plotted by connecting the start and end points with a straight line. From Figure 16 (right) it becomes clear that the route in the left figure is not a good approximation of the real traversed route. The route constructed in right figure is built up from more points, because the site-vector of this IMSI contains more points. A way to get better results for the estimated route is to make use of auxiliary information from for example a planner of motor vehicle routes. Suppose you can assume that length of the traversed route between two points in the site-vector is estimated by the length of the route generated by the planner. This route is realistic in the sense that it is a "real world" route, not just the straight line approximation.

The method of integrating the motor vehicle planner to estimate the distance between two points of the site vector looks at first glance appealing. The question remains if it is possible to integrate the planner in the software we use to estimate the length of a route. Will the computation time remain acceptable for a large number of site-vectors in a sample? Does the extra precision in the estimation of the length of a route compensate for the additional computation time needed to determine the route between two points of the site-vector? Also, we know nothing on the mode of transportation. If the mobile phone user travelled using for example the railroad, the motor vehicle routes would not be correct.