Discussion Paper

Using cell phones to compute dynamic population densities safely: a theoretical exploration

Leon Willenborg

22 February 2025

1 Introduction

The aim of the present paper[1)] is to show how cell phone data can be used to compute information about dynamic population densities, that is densities that change over time because people move. These densities are to be contrasted with densities on the basis of the places of residence of persons, home addresses so to speak. These are more static, as they do not change during a day. They are also more exclusive, as they do not include the persons who do not live at a particular location, but who just happen to be there, temporarily, for instance to shop, for work, to attend a school, or just passing through to another destination, etc.

The densities derived from cell phone data in principle register anyone that is actually at a particular location at a certain moment, under certain conditions: this person should carry a cell phone, have a subscription with a telecom provider[2)], and is actively using the cell phone (i.e. making a phone call, texting, apping (using WhatsApp, Signal, etc.), searching the internet, etc.). So this is quite different from the situation of a person at that location, irrespective of being actively using their cell phone. The question this raises is: how well do such cell phone data produce good proxies of dynamic population densities? To answer this question it would be necessary to consider other sources of personal location data that could be used for estimating dynamic population densities. But such data are likely to come with their own problems when used for this purpose.

This use of telecom data is not new. What might be new is that in the data used by the statistical office all cell phones were anonymous and the data were aggregated. So no cell phone can be tracked. Compare this to e.g. [2] or [5], where individual cell phones are followed over an extended period of time. This required the use of cell phones id's. In contrast, in the model developed in the present paper certain link data are used by the statistical office. These data are prepared by the telecom provider. The link data provide statistical information about (anonymous) cell phones whose presence is spreading over the cells in the network. From this information one can derive the spreading at municipality level. First, population densities for non-overlapping time periods ('hour blocks') are computed. Then, by using the link data, the dynamics of these densities is derived. The approach in the present paper is like studying traffic in a street network based on the movements of anonymous vehicles, where at each junction only the number of vehicles going left, right or straight ahead (etc.) is used. Vehicles are anonymous and therefore they cannot be followed travelling through the network.[3)]

One of the reasons why cell phone data can be useful is that mobile phones are owned and used by many people on a regular basis and incessantly by some. There are, however, groups in the population that typically do not own a cell phone or regularly use one. Any time a cell phone is used for a telephone call, a WhatsApp (or Signal, etc.) message, access to the internet, records are generated by the telecom provider to mark and log this event. It is the basis information to

1) The author would like to express his thanks to several former CBS colleagues for their help with this paper: Lara Reuter skillfully created all figures produced with ArcGIS. Sander Scholtus reviewed several drafts of the paper, which led to significant improvements, both substantively and textually. Finally, Edwin de Jonge, Jan van der Laan and Matthias Offermans provided some comments that helped sharpen certain points raised.

2) Which also happens to supply (intermediate) data to the statistical office.

3) Unless, of couse, there are very few vehicles in the network and far apart, that single vehicles can be followed even if they are anonymous. This may be the case in certain parts of the network and at certain times. But such extremal cases can be dealt with by applying standard statistical disclosure measures for tabular data, if necessary.

bill their clients. But this information can also be used for statistical purposes, adopting certain assumptions. What can be used for this purpose, is the fact that persons carry cell phones around. If a cell phone is actively used at a certain moment in time we know (almost for sure) that this must have been the result of the actions of a person, most likely the user of this cell phone. Then we know that this person at this moment is at the same location as the cell phone, which is near the cell (= antenna) which was then communicating with this cell phone. The geographic locations of the cells are known. Therefore the average location of an active cell phone in some hour block is then also known by the cells through which it was communicating.

The present paper does not treat the theory in the most abstract and general way, on purpose. So the observation period is chosen to be a day. It is subdivided into 24 hour blocks. It is applied to the home country of the author, namely The Netherlands rather than to some abstract country. The idea is to transform certain communication information available on cells to geographical densities, which in our case are defined on municipalities. This requires an intermediate step to achieve. Without further information on the cells and their sensitivity properties it is natural to apply an interpolation technique such as nearest neighbour interpolation which uses Voronoi polygons associated with the cells. These Voronoi densities are then used to compute municipal densities. Municipalities are taken as examples of geographical areas which are statistically relevant, in contrast to the Voronoi polygons, which are only auxiliary constructs.

It should be stressed that the present paper is theoretical in nature, focussing on ideas rather than on empirical work and computations. Because no telephone data were at the disposal of the author, the examples to illustrate certain points are therefore based on fictional data. The locations of cells, used in some examples, are publicly available.

The model used in this paper, though basic, is sufficient to illustrate the essential points of the approach. It is fairly easy to produce variants of the basic model that add realism to it. In such variants one can distinguish cells on the basis of their directional sensitivity (e.g. omnidirectional or unidirectional). And a sensitivity area can be used for each cell that matches its direct environment, which may be affected by the presence of objects such as large buildings, trees, lakes, etc. in its vicinity. This is very specific technical information, which was also not available to the author. In order to avoid describing a rather abstract model with unspecified sensitivity areas per cell, without the possibility of providing some concrete examples to illustrate certain points, the author has chosen to focus attention on a more unsophisticated model to bring certain points across. This seems to be a natural starting point, especially because more realistic features can easily and naturally be introduced to the model subsequently. This topic is briefly discussed in the final appendix of the paper, i.e. Appendix F. But let us first give an overview of what else can be found in the paper.

Section 2 reflects on the very method that is used to estimate dynamic population densities. The method uses active anonymous cell phone data. If a cell phone is active it can be linked to one or more cells during a session, irrespective of whether it is moving or not. This implies that an active cell phone can be linked to the locations of these cells. The fact that cell phone data are anonymous (for the statistical office) means that cell phones cannot be traced over time. Nontraceability of cell phones (and hence their users) implies that the data used by the statistical office are safe, in the sense that they cannot be linked to individuals. The precision of locating cell phones does not exceed the precision of the locations of the cells. Cell phone users are autonomous in deciding to switch on or off their cell phone. Or they may leave or enter the country at any time they want. This makes the population of cell phones an open one. These are some properties that the data possess that we want to use to produce dynamic densities of

persons in the country. There are several more properties of importance or interest. These are discussed in this section, possibilities as well as limitations. The cell density is the basis for all other densities we consider in the present paper: geometric cell densities, Voronoi densities and municipal densities.

In Section 3 there is a discussion about the source data that is used, the output that a statistical office wants to produce from this and the data used are provided by a telecom company. These data are sensitive and cannot be provided directly to the statistical office. Therefore the telecom provider preprocesses the source data in order to produce intermediate data, which are safe enough[4)] and which can be delivered to the statistical office for further statistical use.

In Section 4 cell densities are considered. They are derived from the time being active by cell phones. This produces, for each hour block and per cell, a mass ('total presence'). Because a cell phone can move and because communicating cells may switch even for a stationary cell phone, the presence of a cell phone is typically distributed over several nearby cells (involved in communicating with cell phones) during an hour block. This 'distributed presence' is, technically, not a problem. By adding all these presences for each cell per hour block we obtain a total presence. We shall ultimately not be interested in these absolute presences but in relative ones, that is, their density. This quantity is only auxiliary and serves to compute municipal densities and other similar statistically meaningful geographical densities.

Section 5 is the first step to use the cell densities to obtain geographical densities. In this first step geographical cell densities are derived. A formula is presented that describes this conversion. Section 6 is about special densities that can be derived from the cell densities, namely the Voronoi densities. It also discusses the density mass that can be based on it, which is an important means to translate Voronoi densities to statistically meaningful geographical densities such as municipal densities. The Voronoi densities can also be used directly to provide heatmaps of densities. They should be smoothed first so that the contours of the Voronoi polygons have been masked. Voronoi polygons and Voronoi densities are only semi-finished products, although important ones. They need to be further processed and polished to yield useful finished products. This finishing is discussed in the remaining sections. Section 7 is about the translation of Voronoi densities to municipal densities. This translation can be done in two ways: in a numerical way in which a density value is computed for each municipality, or in a graphical way in which the Voronoi densities are smoothed and presented as heatmaps. The contours of the municipal boundaries are added as a visual aid. They are not part of the computation, as they are in the numerical approach. They form a separate layer in the GIS software, and their only purpose is to give visual assistence. So far the densities have been considered separately from each other. However, a natural next step is to try and understand their dynamics. This can be done on the basis of the movement of cell phones. But with anonymous cell phones this is impossible. However, by letting the telecom provider produce some extra information from the source data, this can be done safely. We start this part of the paper by considering how cells are connected through moving cell phones.

In Section 8 it is shown how to define a digraph structure on top of the set of cells. An arc (a, b) indicates that a moving cell phone active in cell a at some hour block h can be active in cell b at hour blocks h or $h + 1$. So cell b is within reach from cell a within a limited amount of time (that is, in less than 2 hours). The telecom provider can provide these arcs, without the necessity of

4) Certainly against spontaneous recognition, but likely (far) beyond that.

sharing confidential data. They require that the location of cell phones in one hour block as well as in the next hour block is known to produce the basic input. By letting the telecom provider aggregate over the cell phones involved the link data can be derived: If there was at least one cell phone that was active in cell a in hour block h and in cell b in hour block $h+1$, (a,b) is an arc. Which cell phones led to this arc is unimportant, of course, and will not be revealed to the statistical office. Important is that it is possible for a cell phone to be active in cell a in hour block h and in cell b in hour block $h+1$. So this is about a possibility, a potentiality. It is obvious that these link data are safe. However, we need to go further if we want to understand how the changes in population densities come about. As cell phones are anonymous we cannot trace any cell phone. Although we have no complete paths of cell phones we do have information about small parts of these paths, namely how 'density mass' flows among the cells, from each hour block h to its immediate successor hour block $h+1$ (all within the time window considered). The cell flow information that we have is in the form of a sequence of Markov matrices. It is as if the original paths of individuals are replaced by knowledge about the movements of anonymous persons at each junction: it is only known which percentage turns right, or left or moves straight ahead, for each pair of adjacent hour blocks. In fact this means that the original paths are replaced by traffic information, in the form of a nonstationary Markov chain, which, as is well-known, is memoryless. This is the price to be paid to protect the privacy of cell users. This cell density flow problem is studied in Section 9.

In section 10 we consider the geometric cell density flow. This is about the change of cell flow information for the cell locations. From this the Voronoi density flow is derived as well as the municipal density flow. Both of these flows are important end results. The geometric density flow is a (crucial) intermediate result. The Voronoi density flow is discussed in Section 11. It is shown how it can be derived from the cell density flow by a formula that neatly separates this information from geometric information about the Voronoi polygons generated by the cell locations.

In Section 12 we consider the equivalent of the cell link digraph, namely the municipal link digraph. This digraph indicates how density mass flows between municipalities. Using an example it is shown how an adjacency matrix for cell flow is used to compute a strength matrix for clusters of cells (here on the basis of being located in the same municipality) as well as the relative strength matrix. This is the Markov matrix derived from the strength matrix. These matrices provide information about how strong municipalities are connected through travelling cell phones. This connection is overall, not for specific hour block pairs. Section 13 is about the municipal density flow. It is shown how densities for consecutive hour blocks are linked by Markov matrices. Also considered is how Markov matrices for cell flow can be transformed into Markov matrices for municipal flow. In Section 14 the density transformations that have been derived in the present paper have been collected and presented in context. In Section 15 we consider animated density flow, which is a way of showing graphically how densities change over time. We consider Voronoi densities, unsmoothed or smoothed, with municipal contour lines[5)] to visually aid the viewer, presented as heatmaps. The idea is to play these visualized densities in the order of the hour blocks, as a film. It is then immediately clear how a density changes over time.

Now, with all the output densities and their development over time being dealt with, we could

[5)] Of course, in a similar way other geographical partitions can be used, such as COROP areas, provinces and the like. The contour lines ar not part of the computations for heatmaps and therefore can be freely chosen.

stop. However, one step is added, namely one in which the flow in the cell network is analyzed. Section 16 suggests how to do this, namely by using the Helmholtz decomposition. In this way one can separate the flow among cells or municipalities into two components: one measuring the throughflow, the other measuring local flow or circulation.

Section 17 contains a discussion of the main results and some topics that remain open for future research. The main text is concluded with a list of references. A number of appendices complete the paper.

2 Considerations, concepts and caveats

That certain metadata generated by mobile phones is useful data to provide actual/dynamic population densities is not immediately clear. In the present section we discuss several aspects that are involved. And also the assumptions that we apply in order to resolve an issue.

2.1 Selective set of cell phone users

Not everyone uses a mobile phone. This in itself is not necessarily bad news, provided various groups in the population are well (proportionally) represented. But, as a matter of fact, this is not the case. Certain groups are under-represented, such as the elderly or the very young[6)] and hence others are over-represented (the complementery age group). And possibly representation may be an issue with other characteristics as well (such as income, nationality, race,etc.). So adequate representation of the entire population by the group of cell phone owners is not obvious, and, for certain subgroups, it is not even the case. But when would this be problematic? The answer is not so easy to give. In case the mobility pattern of a group is clearly distinct from that of the mainstream cell phone users, one would see differences in population densities, such as is the case with young children, or with sick or very old people that are bedridden or bad on their feet. And people with jobs probably have different mobility patterns compared to those who are unemployed, who also may have less money to purchase cell phones.

2.2 Active cell phones

Even though people may carry a cell phone with them, if it is not used actively in a certain period of time this cell phone does not generate any event record at the telecom provider's system and hence it is 'invisible', unnoticed. It is similar to the situation where the telephone would have been switched off.

'Actively used' should not be construed literally, in the sense that a cell phone user should actually be speaking to or typing at their cell phone. What the phrase means is that a cell phone

6) And very likely the illiterate and the low literates, and persons with certain physical handicaps which do not allow them to operate a cell phone.

was in contact with the network through a cell in the network. This means that messages sent (calls, text messages) are taken into account, whereas messages received are excluded. In Section 3.1 it is explained why this is the case. Also, only the cells used by the telecom provider selected to deliver input data to the statistical office, are taken into account.

So a cell phone is not recorded when it switched off, was not active (for instance when the user was sleeping), was outside the sensitivity area of any cell in the network used by the telecom provider, for instance when being across the border. Also cell phones using WiFi, instead of a cell in the network, to communicate with the internet, are excluded. This concerns a lot of cell phone users, namely those staying at home or who are at their workplace. In cafés, in restaurants, in shopping malls, etc. often WiFi connections are available, which also may be used frequently by visitors. So it is clear that a significant number of cell phone users is not observed, and hence their location is unknown. Possibly all this causes considerable bias in the data.

On the other hand, if a cell phone was used for navigation in a car, it could have been active for a long time, and a large number of cells may be involved.

Details about the communication between users and cells are not always relevant for the method used, as long as any active period of a cell phone can be linked to one or more cells. The data that are used to measure the presence of persons in an area (near a cell) are in fact based on the periods when users were actively using their cell phones, in the sense just described.

A legitimate question is how useful this information is as a proxy to 'being there'. Clearly it is not a good proxy in situations when a cell phone can (typically) not be used as something else requires the user's attention, like driving a vehicle without the possibility of using the cell phone handsfree. And of course, when people sleep they cannot be actively using their cell phone.

One should realize that the method, strictly speaking, measures certain events involving cell phones, which concerns users only indirectly. But what is registered are events triggered by users of cell phones. Whether these are also the registered clients[7)] is not important, unless one would use demographic characteristics of clients.[8)] Then, in some cases, there would be a mismatch between those of the actual user of the cell phone and those of the registered client. But for our application such a possible mismatch is irrelevant.

2.3 SIM cards, telephone subscribers, cell phones and their users

Strictly speaking it is not even the cell phones that are the sources of the events that are of interest in this paper, but the SIM cards inside these phones. These SIM cards should correspond to valid phone numbers to be usable. In many cases a cell phone only has a single SIM card inside. But dual SIM card cell phones exist that can contain two SIM cards.[9)] Typically, persons who have a private phone number as well as a business phone number would be users of dual SIM card cell phones. So instead of using two single SIM card cell phones they use a single dual SIM card cell phone.

7) Usually they probably are, but the clients may, for instance, also be parents paying for the subscriptions of their children.
8) Which we do not in the method proposed.
9) The use of two SIM cards for such cell phones is an option, not a necessity. With only one SIM card inside it is like a cell phone that can only contain a single SIM card.

It is important to reflect on the difference of a single SIM card cell phone and a dual SIM card cell phone in the context of the present paper. It is impossible that a dual SIM card cell phone can be active on both phone numbers at the same time. However, within one hour block such a phone can use its two phone numbers. This implies that a such a phone is likely to be more active and hence the probability that it counts in the statistics is larger. Therefore it is likely to have a bigger presence (see Section 2.4). The use of a dual SIM card cell phone by a user is no different in our application from two single SIM card cell phones carried around by another user.

To avoid these complications it is therefore assumed in the present paper that the number of persons using dual SIM card cell phones[10)] is much smaller than the number of users with a single SIM card cell phone and hence is negligible. Furthermore it is assumed that we can associate each user with a single SIM card cell phone at any point in time. And furthermore that each cell phone is used by a single person, and in particular, is not shared with any other person. The telephone subscribers or the owners of cell phones are irrelevant in the present approach. Important is that the telephone numbers (or the corresponding SIM cards) should be used to count.[11)]

2.4 Presence of cell phones

In this section we deal with a variable called presence, which is used to indicate for a cell phone id_i how much time it communicated with a cell c in hour block h. From the presences of cell phones at cell c in hour block h the total presence is computed. Total presences is the base material for estimating dynamic population densities. The statistical office depends on the telecom provider to prepare total presence. Presence of a cell phone is considered sensitive intermediate data that should be available to the telecom provider only. The statistical office will only receive this information in aggregated form, that is, as total presence per cell and per hour block.

We consider two possible definitions of presence for cell phone id_i in hour block h that was active at cell phone c for some time:

- b-presence, which indicates whether or not cell phone id_i was active at cell c in hour block h.
- q-presence, which indicates which fraction of hour block h cell phone id_i was active at cell c.

b-presence is considered in Section 2.4.1 and q-presence in Section 2.4.2. It should be noted that q-presence is at the the time of writing a quantity that is not (readily) available, as it is not important for the billing of customers to record how much time cell phones have spent communicating through particular cells; only the total time each cell phone used is what matters here. But maybe in the future this information will be (readily) available. At the moment, q-presence can probably only be estimated reliably, using statistical information about the actual switching behaviour of cells. However, even if q-presence cannot be used at the moment, b-presence can be used; q-presence remains a beckoning prospect. Because q-presence is the variable of choice rather than its surrogate, b-presence, it will be used in the remainder of this paper to quantify presence.

10) Or who carry (and use) two (or more) cell phones.

11) In case subscribers would be counted one runs into problems in case a person has two subscriptions, one for him or herself and one for their child. For then there is a possibility that this person is bilocated (present at clearly different locations) at the same or nearly the same time, which is nonsense and has to be avoided.

2.4.1 b-presence

In this section we assume that the variable 'presence' is a binary variable: a cell phone has either been active at a particular cell in a certain hour block, or not. How long the cell phone was active at a cell in that hour block is not known. Present at a cell or not is of importance, which is a binary property. Hence binary presence or b-presence. This in contrast to the concept that measures the amount of presence, namely quantitative presence, or q-presence. This is dealt with in Section 2.4.2.

So for each cell phone id_i it should be established in which hour blocks h (in the observation window) it was active. Then for each cell phone id_i it should be established, for each of its 'active' hour blocks, at which cells c it was active. From this information it can be derived for each cell phone id_i and each hour block h (in the observation window) at how many cells it was active. Also it should be derived from the telecom data (by the telecom provider) how much time cell phone id_i was active in each hour block h, or rather, what fraction of h. This is total time, not the time per active cell.

So recapitulating we know, for cell phone id_i, the fraction of the time $f_{i,h}$ it was active in hour block h and also the number $k_{i,h}$ of cells that were used to communicate with the network. Without any extra details about the communication we may assume that the time spent at each active cell was equal, which is $1/k_{i,h}$ for each of the active cells involved. So an estimate of the presence for each of these cells is $f_{i,h}/k_{i,h}$. The total b-presence $p^b(c,h)$ of a cell c in hour block h is now defined as the sum of the b-presences of all the cell phones id_i active in cell c in hour block h, that is

$$p^b(c,h) = \sum_i f_{i,h}/k_{i,h}, \tag{1}$$

where i indexes the cell phones active in hour block h at cell c.

2.4.2 q-presence

If we were able to step up to q-presence from b-presence we would have the time $t^i_{c,h}$ spent by cell phone id_i at a cell c in hour block h, which in the case of b-presence would have to be estimated. We define the q-presence of a cell phone id_i active at cell c in hour block h as the fraction of one hour that it was active, that is,

$$p_{hci} = t^i_{ch}/60. \tag{2}$$

Obviously, $0 \leq p_{hci} \leq 1$.[12] We define the total q-presence of a cell c in hour block h as the sum of the q-presences of all the cell phones active in cell c in hour block h:

$$p^q(c,h) = \sum_i p_{hci}, \tag{3}$$

[12] We define $p_{hci} = 0$ for cell c in hour block h if cell c was not active for id_i in hour block h.

2.5 Open population of active cell phones

The fact that the persons using a cell phone can switch them on or off, or can come into the country, or leave it whenever they want, creates a problem that can be described as: lack of mass preservation. Mass (i.e. the number of active cell phones at a certain point in time) fluctuates over time. It is not the case that cell phones (or their users) appear or disappear from the universe, but not being recorded by a cell at one hour block does not preclude the possibility of being recorded at the later time. Likewise, being active on the phone at one time does not preclude inactivity at a later time.

Not only this randomly appearing or disappearing of cell phones is typical for cell phone data, also the time being active on cell phones differs among persons and also may differ for the same person over time. This implies that the total presence, summed over all cell phones, is likely to fluctuate for the various hour blocks. There also is not necessarily preservation of q-presence for individual cell phone users over time, let alone for the population of cell phone users. This quantity is likely to fluctuate.

For this reason we have chosen to look at population densities based on q-presence, rather than on total q-presence. Densities have the advantage that they are normalized to 1, whereas total q-presence is not and hence fluctuates over time. The idea is that by using densities, the problem with cell phones becoming active or inactive (at random moments) is (apparently) sidestepped. If one would use total presence then one would have to model the appearance or disappearance of cell phones per hour block and perhaps for (certain) cells as well. This can undoubtedly be done, but it is more complicated than the approach we have opted for in this paper. It is of interest to investigate which effect this moving in and out of the population has on the cell density estimates. See Sections 4.2 and 9.3 for additional information on this topic.

2.6 Cells and Voronoi polygons

We can represent cells as points in a map. We want to derive map information from the measurements of active cell phones in relation to cells. By applying a special interpolation technique, called nearest-neighbour interpolation (see Appendix B , Section B.1), we can achieve this. This technique generates areas, called Voronoi polygons, around each cell, with constant density values. These Voronoi polygons form a partition of the country L (The Netherlands).[13)] A Voronoi polygon linked to a cell c consists of the points in the plane (or in L) that are closest to c. These polygons are used to define densities, Voronoi densities, which are crucial for deriving municipal densities and smoothed densities (by using natural neighbour interpolation; see Section B.2 in Appendix B).

It is tempting to consider the Voronoi polygons as sensitivity areas of the corresponding cells. A sensitivity area of a cell is an area in which the cell can pick up cell phones signals. But Voronoi polygons are only sensitivity areas by crude approximation. The reason is that the sensitivity of cells, which are in fact antennas, depend on a lot of parameters, relating directly to the cells or to the environment in which they are positioned. Determining the sensitivity area is a complicated, technical problem, one that we want to bypass in this paper. The important role of Voronoi

[13)] In the computational geometry literature this is often referred to as ‘Voronoi tessellation’. We shall not use this term in the present paper.

polygons is that they are used to translate cell density values to geometric densities. They are used for smoothing.[14)]

Remark It is actually impossible to model the sensitivity areas without additional information from the telecom provider about each of the cells. They are not all of the same type: some are omnidirectional antennas, which are equally sensitive in all directions, whereas others are unidirectional antennas, which are more sensitive in a particular direction. Furthermore, external conditions are also important for the reception of signals, such as the environment where the cells are located: the presence of high buidings, trees, lakes and other large bodies of water, etc. Also, atmospheric conditions also determine the quality of the cell phone signals received by a cell. And these are not constant but change over time. It is clear that modelling the reception area of a cell in a realistic way is a complicated matter (requiring knowledge of antennas) and it also requires a lot of additional information that cannot be expected to be available to the statistical office. See for instance [2] or [5] for a bit more background to the technical aspects. This sort of information is not of the kind that is typical for a statistical office. As a very crude approximation of a sensitivity area of a cell we can take its Voronoi polygon. But we should keep in mind that the real use of these polygons is a computational one, namely as a means to transform densities of one kind to that of another kind. The Voronoi polygons form a neat partition of the country, so they do not overlap. The sensitivity areas at best form a coverage of the country, but are likely to overlap. So they do not form a neat partition of the country. For all these reasons we leave out realistic sensitivity areas in the present paper and consider Voronoi polygons instead. □

There is a problem when using a Voronoi partition generated by the cell locations: the cells at the perifery of the country ('boundary cells'). Without extra actions (i.e. clipping / truncating considering boundary cells across the country border) they may be unbounded, so that they have infinite size, and are hence useless. In Section 2.7 we look more closely at this issue.

2.7 Interior and boundary cells

If we consider the cell network in country L and we only have information on these cells, we are forced to distinguish between two types of cells based on their location: near the country border or not. In the former case we are talking about boundary cells and in the latter case about interior cells. The problem with boundary cells is that they are unbounded. In particular their area is not defined (∞ in size), which is a complication as we use these areas as weights in calculations.

If information would be available, it would be possible to extend the collection of cells in the network of country L with extra cells: terrestrial cells near the borders with Belgium and Germany and offshore cells, in the North Sea near the coast of L or in large bodies of water such as the IJsselmeer.[15)] Then the cells in L would all be interior cells, so to speak, and hence they would all be bounded. However, even then this would not solve all problems with boundary cells (in L). We still need to know for these cells which part (in terms of area) of these polygons is in L

[14)] We assume that it is reasonable to distribute a cell density uniformly over the corresponding Voronoi polygon. However with the appropriate cell information available we could apply nonuniform densities of the polygons. But we do not have this information. Clearly using it would make computations, for instance for municipal densities, more complicated.

[15)] The offshore cells and the terrestrial boundary cells just across the border in Germany or Belgium are to be treated separately, as external boundary cells.

(in case of terrestrial boundary cells near the German of Belgium border) or which part (in terms of area) is covered by land, in case of coastal cells. We shall not anticipate the deliberations that motivate the choices for truncated or complete Voronoi polygons. This is reserved for Section 6.5, which is entirely devoted to these matters.

The choice between truncated or the original untruncated Voronoi polygons is important for the weights to be associated with a cell. These weights play a role when q-presence is used to compute municipal densities per hour block. See Section 7 for details.

2.8 Some information about the distribution of cells

We present some initial information about the cells in The Netherlands (in January 2023). There were then a total of ± 51300 cells. The minimum number of cells in a municipality was 9 (in Rozendaal) and the maximum number was 2086 (in Amsterdam). The average number of cells per municipality was ± 150. In Figure 2.1 the distribution of the number of cells per municipality is shown. The Pareto graphs provides information on the inequality of the contributions to the distribution. About 90% is less than the average value.

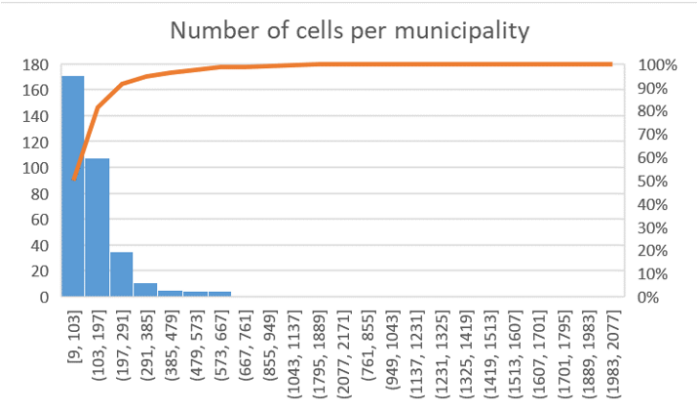

Figure 2.1 Distribution of the number of cells per municipality (in January 2023) and corresponding Pareto graph.

In Figure 2.2 the distribution of the number of inhabitants in a municipality per cell in that municipality is presented. As the graph shows the highest and lowest fraction differ by a factor around 8, with an average of about 400. This gives a crude idea how well inhabitants are served by cells. Of course, the sensitivity areas of cells do not respect the boundaries of municipalities. And cells are not only used by the inhabitants of the municipalities where they are located.

More information about the geographical distribution of cells (on land and in the waters near the Dutch coast) can be found in the figures in section 6.4.

2.9 Restrictions on the use of anonymous cell phone data

We now want to compare the possibilities of anonymous, untracked cell phones (as in the present paper) with the possibilities of anonymous, tracked cell phones (for some period of time)

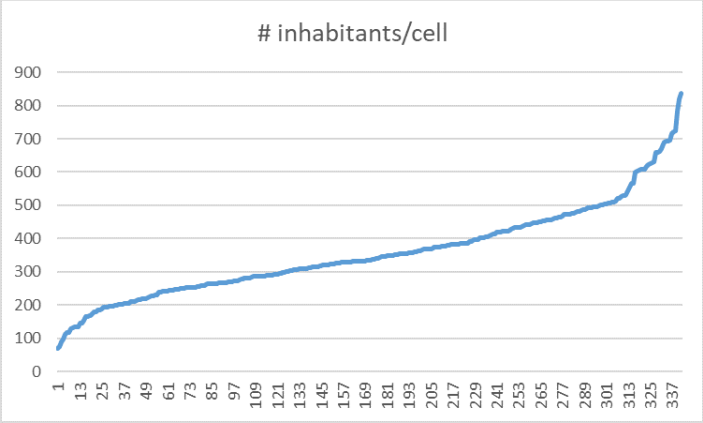

Figure 2.2 Distribution of the number of inhabitants of a municipality for each cell located in this municipality (in January 2023).

as considered in [2] or [5].[16)] Anonymous means that no id's (keys) of cell phones are used nor surrogate keys, to allow tracking of otherwise unidentified cell phones. In particular we focus on some limitations of the method used in the present paper compared to the approaches in [2] or [5]. As remarked before, it should be borne in mind that the methods have in common that the locations are those of the cells; there is no more detailed location information available in the data.

In the latter case it is possible to determine the location (cell) where the cell phone has spent most of its time during the tracking period. In that way one is able to have a good guess for some users where their abode is (at the cell level). These users are local residents or tourists staying at a hotel or guesthouse, provided they stay there long enough. In case of workers who work regularly at a fixed location (say a factory or an office) one would perhaps also identify their 'place of work'. This depends on the period when these persons are tracked and the length of the tracking period. Persons who work mainly from home would be undistinguishable from persons who are not working at all, because they are unemployed, pensioned or ill at home or in a hospital, daycare, nursing home, etc. (during the observation period). Such results are beyond the reach of the kind of data the present paper proposes to use. It is even impossible to measure the presence well at what is the home locations for most people, namely where they sleep. Sleeping implies nonactivity of users on their cell phones and hence that their presence is not registered by any cell.

The view we get of the dynamics of the population by using the link data provided by the telecom provider is that of a time dependent Markov chain. This means that there is no memory in such a system. The current state summarizes past and present. In reality, people travelling is typically not memoryless but purposeful. People tend to make trips with plans. They leave the house in the morning and come back home in the evening. The same with tourists. Replacing all these motivated and purposeful trips at the micro-level by a purposeless drifting around following the rules of a Markov chain helps to make the data safe without giving up the possibility to study the movements of an anonymous collective of people. This is typical for most

[16)] In these papers the id's of the cell phones are replaced by surrogate keys to allow tracking the phones. A surrogate key has in common with a key that it uniquely links to individuals (in this case clients of the telecom company), however without knowing who they actually are. It can be used to link information about the same individuals spread over time. It is important that information about the same individuals can be combined, not who these individuals actually are. The use of surrogate keys gives a degree of protection, but not necessarily against recognition of individuals on the basis of unique routes at the level of cells.

traffic studies. To study how crowded roads, etc. are at particular times, there is no need to use details of the travellers, including the purposes of their trips, their destinations, where they started their trips, where they live, etc.

3 Description of various data sets

Part of the data generated by the telecom system is used by providers to bill their subscribers for using their cell phones. This billing information is also useful as input for statistical offices wanting to estimate the dynamic population density over some period of time. In the present section it is discussed how this basic information owned by telecom providers can be used safely, that is without compromising the privacy of subscribers or customers. The telecom provider should process the original data, which are not safe, to obtain intermediate data, which are safe, that can be delivered to the statistical office.

The present section starts considering the source data, and some of its characteristics, in the light of the proposed application. It also draws attention to properties of these data that are not ideal. Also the intermediate data are discussed, in particular the intended application is in mind, as well as safety issues.

3.1 Source data (telephone subscribers → telecom provider)

We assume for simplicity that there is a single telecom company providing the data to the statistical office for producing information on dynamic population densities. In practice there may be several such companies. Each of them is supposed to produce the same kind of data as described here. Differences being eliminated they may be considered as being provided by a single (virtual) telecom provider. To produce such seamlessly integrated data may be less straightforward than it looks on paper and it may be time consuming. However, we are not interested in the details of such a process in the present paper.

In order to bill its clients the telecom provider collects information about their use of their cell phones: to make phone calls, to send text messages, to browse the internet, to upload or download data, etc. Not only the events are recorded but also the time a phone call lasted, how much data was transferred and which cells were used when these events took place. This latter piece of information is important for the charging of clients as they may pay a different tarriff when being abroad compared to their home country. This information is used by the telecom provider not only to bill the clients but also to verify to which extent they have used their resources and to compute what resources are left for the rest of the month.

Remark This billable information is sufficient in case users search or browse the internet or send a message to another cell phone user (via e-mail, WhatsApp, Signal, etc.). There either is no receiving cell phone involved (in case of searching or browsing) or there is, but receiving the

message, takes a very short time and there is no subsequent action.[17] It shows that there is an asymmetry in handling messages between senders and receivers. It therefore seems inappropriate to measure presence at the receiving end. Only in case the receiver replies to a telephone call, this would seem appropriate. In case the receiver replies to a text message he or she is an active user in the sense described above and presence can be measured. Therefore it seems appropriate and not restrictive that only billable information is used. □

It should be noted that the client is not necessarily the user of the cell phone. Think of a parent paying for the cell phone of their son or daughter. But for our purposes it is relevant that, with an active cell phone, one person is linked and with different cell phones different persons. See Section 2.3.

3.2 Basic data (telecom provider)

We assume the basic data file to consist of records with event information, where an event is as discussed above. It is supposed to contain (at least) the following variables:

- user-id (id of a client)
- event-id (to combine records starting and ending the same event so the duration of the event can be computed)
- start/stop (to indicate whether the event started or ended)
- cell-id (for the cell (=telephone antenna) that was active when the event started or ended)
- timestamp (a detailed date and time indicator to mark the start or end of an event)

The records may contain more information, but this is what we need for our purposes. These variables allow us to make the necessary selections, groupings and links that we need.

Because the source data are sensitive, the source data have to be processed by the telecom provider who produces safe, intermediate data. These data will be delivered by telecom provider to the statistical office, who wil use these data to compute the results they are interested in. The kind of output the statistical office would like to make is described in Section 3.4. How the telecom provider produces these intermediate data from the source data is discussed in Section 8.

3.3 Intermediate data (telecom provider → statistical office)

Cell count data per hour block provide the basis for population densities per hour block. We could leave it at that: a sequence of population densities. But then we would not understand how the 'population mass' moves across the cell phones. As we use data of anonymous cell phones we are not able to follow any cell phone over time. Between these extremes there is an intermediate solution: link data that link cells, in the sense that the transition between cells is

[17] An e-mail, for instance, is received by a mail server and the person to whom it is addressed may not be immediately informed when it is received and he or she may look at it (much) later. This is a case of asynchronous communication. In case of a WhatsApp (or Signal, etc.) message there is a notification to the receiver, but, again, he or she may look it up later. This shows that there is a big difference between sending a text message and receiving it as far as the activities of sender and reciver. This is quite different with a telephone call, when both sides of the messaging are talking, when a link is established. This is typical for synchronous communication.

given, due to moving cell phones. These link data are to be provided by the telecom provider to the statistical office so that it can do computations concerning the dynamics of the population densities from one hour block to the next.

The intermediate data should be safe, in the sense that no information about specific persons can be inferred from the data. This can be done by producing certain aggregates from the individual cell phone data. These aggregates should be produced per hour block. For these aggregates the 'true identity' of cell phones is not important. It is only important that they have some kind of temporary id (that is kept within the premises of the telecom provider), so that can be decided, per hour block which activity data are produced by the same phones. Information in different hour blocks is treated independently of that of other hour blocks. In fact, the same cell phones producing data in different pairs of hour blocks could have different temporary id's, as they are not linked across different hour blocks. These temporary id's are only used by the telecom provider when compiling the intermediate data as auxiliary variables. They do not appear in the intermediate data as delivered to the statistical office.

The intermediate data is supposed to contain the following information:

- Information to derive cell densities. This information is total presence, which is derived from all the active cell phones at a particular cell c in a particular hour block h.
- information linking cells in consecutive hour blocks through moving and active cell phones. This information can be supposed to be given in the form of a set of $(0,1)$-matrices (providing information about which cells are mutually connected in the sense just described) and by a set of Markov matrices (indicating the probability of a cell phone being active at cell i at hour block h is also active at cell j at hour block $h+1$.

3.4 Output data (statistical office → general public)

Here we give an overview of the kind of output the statistical office intends to produce on the basis of cell phone data. In later sections we go into the details of the work to be done, which is partly done by the telecom provider and partly by the statistical office.

The reason for this division of work is that the original data are considered confidential and therefore has to stay within the premises of the telecom company. By delivering certain pre-processed data the telecom provider can make sure that the data delivered to the statistical office are safe on the one hand and flexible enough for the statistical office to produce interesting results about the dynamic population densities. We call the data delivered by the telecom provider intermediate data. More about these data can be found in Section 8.

The first thing the statistical office would want to publish is the densities per hour block. This involves several issues. The main one is about estimating the population densities at the cell level. The next one is to translate the results at the cell level to the municipality level. It involves interpolation and smoothing of population densities, among other things. In Section 7 these topics are studied in detail.

How the population densities change from hour block to hour block is the next topic considered. See Section 14. This topic is focused at understanding how cell densities for subsequent hour blocks change due to movements of cell phones. The direct way to study the dynamics of the

densities is to consider the cells. We have to take into account that we are dealing with an open population of cell phones: inactive cell phones can be activated, and activated ones can be deactivated, at any moment. Likewise cell phones from abroad can enter the country, or leave it. And, of course, new cell phones can be bought and used, while existing ones can be lost or irrepairably damaged. Such abrupt changes are to be viewed as random events. But the study of the change of densities for a statistically meaningful geographical partitioning (such as municipalities) is also of interest. Both are considered in this section.

4 Cell densities

This section describes the cell count data that the telecom provider is supposed to derive from the source data and deliver to the statistical office. The source data with data about the use of the services provided by the telecom provider to its clients are aggregated in such a way that the resulting data can be used by the statistical office to compute population densities per hour block. The reason that the telecom provider should carry out this aggregation is the sensitivity of the base data. The resulting aggregate data are safe and can therefore without any problem be delivered to the statistical office.

If the time intervals were very small each cell phone would be allocated to one cell when active. But we are dealing with time intervals of a length of one hour and moving, active cell phones can be picked up by several cells in such a period.

Table 4.1 is a subtable of Table A.1, where the id's of the cell phones have been removed as they are not important when determining the total q-presence of active cells in the respective hour blocks. The record numbers of the original table have been retained for easy reference, although they are not used here.

We can summarize Table 4.1 in Table 4.2, with the total presence per cell for each hour block.

As we have a total of seven cells in our toy example, we can compute the population density for each hour block h, $h+1$ and $h+2$ from Table 4.2. The results are presented in Table 4.3.

Results as in Table 4.3 are basic for the approach in this paper. They give densities over the collection of cells per hour block. They will be used to compute densities for geographic areas such as municipalities. First they will be used to compute Voronoi densities, which are based on Voronoi partitions induced by the location of the cells (see Section 6). Then, in turn, these Voronoi densities are used to compute population densities for standard geographical subdivisions such as municipalities (see Section 7). This completes the static part of the density problem.

4.1 Cell density defined

A formal definition of a cell density f_c^h, where the subscript c denotes that we are dealing with a density on cells and h denotes the hour block, is

rec	h bl	cell	q-pr	rec	h bl	cell	q-pr
1	h	c_1	0.67	21	h+1	c_6	0.78
2	h	c_2	0.19	22	h+1	c_6	0.15
3	h	c_2	0.08	23	h+1	c_7	0.48
4	h	c_3	0.25	24	h+2	c_2	0.38
5	h	c_3	0.33	25	h+2	c_2	0.21
6	h	c_4	0.21	26	h+2	c_3	0.19
7	h	c_4	0.14	27	h+2	c_4	0.41
8	h	c_5	0.13	28	h+2	c_5	0.14
9	h	c_6	0.21	29	h+2	c_5	0.12
10	h	c_6	0.49	30	h+2	c_7	0.13
11	h	c_6	0.16	31	h+2	c_7	0.63
12	h	c_7	0.71	32	h+2	c_1	0.26
13	h+1	c_1	0.18	33	h+2	c_1	0.47
14	h+1	c_2	0.63	34	h+2	c_2	0.19
15	h+1	c_3	0.72	35	h+2	c_3	0.22
16	h+1	c_3	0.23	36	h+2	c_4	0.23
17	h+1	c_4	0.29	37	h+2	c_5	0.11
18	h+1	c_4	0.13	38	h+2	c_6	0.16
19	h+1	c_4	0.24	39	h+2	c_7	0.15
20	h+1	c_6	0.11				

Table 4.1 Records of active cell phones (without id) at cells (cell) in hour blocks (h bl) h, $h+1$ and $h+2$ and q-presence per cell phone (q-pr).

cell	h	$h+1$	$h+2$
c1	0.67	0.18	0.73
c2	0.27	0.63	0.78
c3	0.58	0.95	0.41
c4	0.35	0.66	0.64
c5	0.13	0	0.37
c6	0.86	1.04	0.16
c7	0.71	0.48	0.91

Table 4.2 Total q-presence at each of the seven cells for the hour blocks h, $h+1$, $h+2$.

$$f_c^h = \frac{p^q(c,h)}{\sum_c p^q(c,h)}, \tag{4}$$

where $p^q(c,h)$ is the q-presence as defined in (3) and the sum in the denominator is over all the active cells in hour block h, that is, with $p^q(c,h) > 0$.[18)]

It should be stressed that at this stage the cells are without locations. They only have an identity so that they can be individually distinguished. Let $\mathcal{C} = \{c_1, \dots, c_n\}$ denote the set of n cells. We now define

$$f_c^h : \mathcal{C} \to \mathbb{R}\backslash\mathbb{R}^-, \tag{5}$$

[18)] In case b-presence is preferred one can replace $p^q(c,h)$ by $p^b(c,h)$ as defined in (1) and one obtains a different definition of a cell density.

cell	h	$h+1$	$h+2$
c1	0.188	0.046	0.183
c2	0.076	0.16	0.195
c3	0.162	0.241	0.103
c4	0.098	0.168	0.16
c5	0.036	0	0.093
c6	0.241	0.264	0.04
c7	0.199	0.122	0.228

Table 4.3 Population density for the hour blocks h, $h+1$, $h+2$ at the cell level.

where $f_{c_i}^h$ is the cell density value for hour block h at cell c_i, with

$$\sum_{i=1}^{n} f_{c_i}^h = 1. \tag{6}$$

It turns out to be handy at several occasions in the sequel to define the column vector with density values for hour block $h = 1, \ldots, 24$:

$$f_C^h = \left(f_{c_1}^h, \ldots, f_{c_n}^h\right)'. \tag{7}$$

Define the all ones column vector ι_n of length n as

$$\iota_n = (1, \ldots, 1)' \in \mathbb{R}^n, \tag{8}$$

then (6) can be written as

$$\iota_n' f_C^h = 1 \tag{9}$$

Quite a large part of the paper is about (5), where the actual location of the cells is not important, only the (density) mass assigned to each cell. This concerns the cell densities per hour block as well as the cell density flow.

4.2 Assigning cells to inactive cell phones

So far we treated q-presence less than 1 as a fact we had to live with. This is not true, however. As stated in Section 2.4.2 cell phones that are not active have not vanished from the earth; they still are somewhere. If they would become active (within the range of the cell network) they would be linked to a cell. On the basis of available data we make reasonable assumptions where it may be. We distinguish two types of cells. Interior cells, which are too far removed from the border for a cell phone to cross the border. The only option for cell phones near such cells is that they can be switched on or off. Cell phones near the border have two options when they

suddenly apppear or disappear: they were switched on/off, or they crossed the border. This latter phenomenon can only occur in certain cells, namely those on the border (with Belgium, Germany or in the North Sea, near the Dutch coast).

Of course, the statistical office does not know what exactly happened with cell phones near the border on the basis of the data provided by the telecom provider, which are far more limited than the original data collected by the telecom provider: were they switched off or did they cross the border?

Suppose we[19] want to impute incomplete observations, that is, for cell phones for which the presence p in an hour block h is strictly less than 1. For the remaining of the presence $1-p$ we want to impute cells where it would be located if the cell phone would be active. Suppose that in the presence part of the cell phone it was associated with cells $c_1, \dots, c_k$. Suppose that the presence for these cells in hour block h is $p_1, \dots, p_k$. with $p = p_1 + \cdots + p_k < 1$. We assume that when no presence was recorded the cell phone was in the vicinity of $c_1, \dots, c_k$ and with presence $p_i' = p_i/p$, for $i = 1, \dots, k$, in which case $p' = p_1' + \cdots + p_k' = 1$.

Of course, this is a rather simplistic imputation model, but it is one that can be used as an alternative for dealing with the missing data by ignoring them. So using this model we would obtain two different estimates for total presence (see Section 4) and for Markov matrices that describe transitions between cells for neighbouring hour blocks, like h and $h+1$ (see Section 9).

5 Geographical cell densities

The geographical cell densities that we want to consider here, have two components: cell locations $\ell(c_i)$ of cell c_i and the cell density value $f_{c_i}^h$ at this cell in hour block h. So we have the cell density value of each cell in hour block h defined on a location (on a map). This offers the possibility for two things that are important in the sequel: we can smooth these densities and get a Voronoi density on a map, where the cell locations are used as generators of the Voronoi partition. The Voronoi density can be used to compute a municipal density. And it can be used to smooth it geographically and compute a density that can be visualised. A geographical cell density also allows us to group the cells on the basis of location in the same municipality. This in turn allows us to compute the municipal density flow using the cell flow. Clearly geographical densities play a key role in computing some important output data.

For presenting the output it is of importance to use the locations of the cells. So that is our next step.

In the next step we assign a location (in $\mathbb{R}^2$) to each cell, and to each location a density. This is a prelude to a type of density in the plane that is central to this paper, namely a Voronoi density. See Section 6. Voronoi densities are used for two purposes: to help define densities on geographically and statistically meaningful partitions such as municipalities. And on the other hand to produce smoothed densities that can be displayed on a monitor or printed on paper.

[19] A desire of the statistical office to be carried out by the telecom provider

So let

$$\ell : \mathcal{C} \to L \subseteq \mathbb{R}^2 \tag{10}$$

be the function that assigns a location to each cell in $\mathcal{C}$ to a spot in L. The function ℓ is injective, as two different cells are located in two different positions.

We now define a density function in the plane where all the mass is concentrated in some points, namely the locations of the cells, where at the location of cell c_i there is a mass $f_{c_i}^h$. We can present the geographical cell (GC) density as

$$f_{GC}^h = \sum_{i=1}^{n} f_{c_i}^h \mathbb{1}_{\ell(c_i)}, \tag{11}$$

where $x \in \mathbb{R}^2$, f_c^h as in (5) and $\mathbb{1}_z$ is the indicator function for $z \in \mathbb{R}^2$.

$$\mathbb{1}_z(x) = \begin{cases} 1 \text{ , if } x = z, \\ 0 \text{ , if } x \neq z. \end{cases} \tag{12}$$

We will rewrite (11) in a form that is more suitable for our application. To do this we first need to introduce some notation. We first define the following column vector of indicator functions:

$$\vec{\mathbb{1}}_{GC} = \left(\mathbb{1}_{\ell(c_1)}, \dots, \mathbb{1}_{\ell(c_n)}\right). \tag{13}$$

Now we can write (11) as

$$f_{GC}^h = \vec{\mathbb{1}}_{GC} \cdot f_C^h = \mathcal{G} f_C^h, \tag{14}$$

where 'cdot' denotes the standard inner product in $\mathbb{R}^n$ and f_C^h is as defined in (7). $\mathcal{G}$ is a linear transformation.

6 Voronoi densities and mass

In the present section we start with considering Voronoi partitions generated by (the locations of) the cells. The Voronoi polygons act as intermediates between cell information and geographic information.[20] Cells on a map can be viewed as dots and Voronoi polygons are elementary

[20] They can also be viewed as proxies of the sensitivity areas of the cells, if one wishes, although this simile should not be taken too seriously.

pieces of area, that form the linking pins to statistically more meaningful geographic areas such as municipalities. Cells and maps match $1:1$. The cell (location) can be viewed as its centerpoint of the corresponding Voronoi polygon. We can translate cell densities to Voronoi densities, which in turn can be used to deduce municipal densities (see Section 7). These densities are defined per hour block h.

6.1 Voronoi partitions

Let $p_1, \ldots, p_n$ be a set of distinct points in the plane. We assume the Euclidean distance $d_2(\cdot,\cdot)$ to be used in $\mathbb{R}^2$ which is derived from the norm $\|\cdot\|$:

$$d_2(x,y) = \|x-y\| = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}. \tag{15}$$

For each p_i we define a region in the plane

$$V(p_i) = \{x \mid \|p_i - x\| \leq \|p_j - x\| \text{ for } j \neq i \text{ and } i,j \in N_n\}, \tag{16}$$

where $N_n = \{1, \ldots, n\}$. $V(p_i)$ is the Voronoi polygon associated with p_i, $i = 1, \ldots, n$. The p_i are the generators of this Voronoi partition. In this paper the generators are the cell locations. The Voronoi partition is the result of the application of nearest neighbour interpolation to the set of locations of cells. This method is discussed in Section B.1 of Appendix B.

Figure 6.1 shows an example of a Voronoi partitioning in the plane. For each Voronoi polygon its generator as indicated as a dot. The Voronoi polygons in the center are presented in full. Those at the boundary are truncated (clipped).

Note that the choice of the metric (d_2 in case of (16)) also determines the shape of the polygons that are generated in the partitioning, although this is not always stressed, for obvious reasons: the metric on the set considered (L in this paper) is often fixed, as it is in the present paper.

For more information on Voronoi partitions the interested reader is referred to [8], where they are called Voronoi tessellations.

6.2 Voronoi densities

In the present section we show how a geographical cell density can be transformed into a density defined on the corresponding Voronoi partitioning, with the set of cells as generators. This transformation is obtained by the application of nearest-neighbour interpolation (see Section B.1 in Appendix B).

Let f_{GC}^h as in (11) denote the geometric cell density for hour block h. This density is 0 in most places, except at the locations of the cells. For each cell c_i the value is $f_{c_i}^h$, which is the total q-presence measured at cell c_i in hour block h. If we apply nearest-neighbour interpolation to this function what happens is that the 'density mass' concentrated at each cell location $\ell(c)$ of

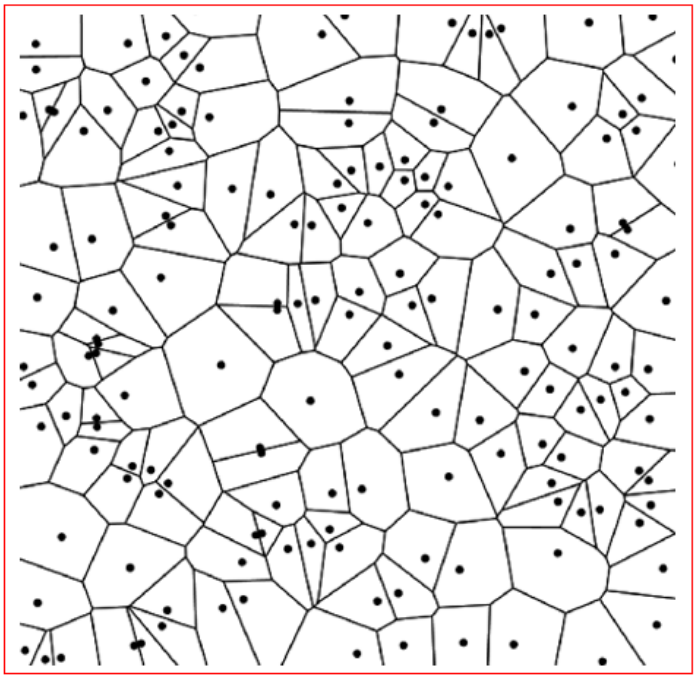

Figure 6.1 An example of a Voronoi partition in the plane.

cell c is uniformly spread over the area consisting of points that are closest to this cell c rather than to any of the other cells, that is, over the Voronoi polygon $V(\ell(c))$ of cell c. Let the area of a Voronoi polygon V be denoted by $|V|$. What we actually want to do is to allocate the original 'density mass' associated with cell c and concentrate it in location $\ell(c)$. This mass is supposed to be evenly distributed over $V(\ell(c))$. The resulting Voronoi density for hour block h, f_V^h, can be expressed as:

$$f_V^h = \sum_{i=1}^{n} \frac{f_{c_i}^h}{|V_i|} \mathbb{1}_{V_i}. \tag{17}$$

where $f_V^h : L \to [0,1]$, with $L \subset \mathbb{R}^2$ representing The Netherlands on a map (in case of the present paper), and where we have written V_i instead of $V(\ell(c_i))$. Also $\mathbb{1}_{V_i}$ is a generalization of the indicator function $\mathbb{1}_z(x)$ defined in (12), where

$$\mathbb{1}_{V_i}(x) = \begin{cases} 1 \text{ , if } x \in V_i, \\ 0 \text{ , if } x \notin V_i. \end{cases} \tag{18}$$

Now we want to rewrite (17) in a similar way as f_{GC}^h as defined in (14). We first define the following indicator function for Voronoi polygons which is like (13):

$$\mathfrak{v} = \left(\frac{1}{|V_1|} \mathbb{1}_{V_1}, \cdots, \frac{1}{|V_n|} \mathbb{1}_{V_n} \right). \tag{19}$$

We now can express (17) as

$$f_V^h = \mathbf{v} \cdot f_C^h = \mathcal{V} f_C^h, \tag{20}$$

where f_C^h is defined in (7), '$\cdot$' denotes the standard inner product in $\mathbb{R}^n$ and $\mathcal{V}$ is a linear transformation, independent of h. The latter property is true as long as the network of cells does not change. This is what we assume in the present paper.

6.3 Mass of a Voronoi polygon and subsets thereof

In addition to the Voronoi density f_V^h for hour block h we also need the total mass associated with Voronoi polygons, or parts of it, in our applications. This includes the treatment of boundary cells as well as the derivation of municipal densities. In case of a boundary cell $\bar{c}$ with Voronoi polygon $V_{\bar{c}}$ we may have to consider the truncated Voronoi polygon $V_{\bar{c}} \cap L$ instead of $V_{\bar{c}}$. In case of transforming Voronoi densities to municipal densities we also have to deal with truncated Voronoi polygons. This is the case when a Voronoi polygon V is incident with a municipality M, that is, when $M \cap V \neq \emptyset$.

Let a (density) mass f_c^h be associated with cell c in hour block h be given.[21] Let V_c denote the Voronoi polygon associated with cell c. The density mass f_c^h is assumed to be uniformly distributed over V_c. This yields a density $\bar{f}_c^h$, which equals the density mass f_c^h uniformly distributed over V_c:

$$\bar{f}_c^h = \frac{f_c^h}{|V_c|} = \frac{m^h(V_c)}{|V_c|}, \tag{21}$$

where $|V_c|$ denotes the area of V_c. $m^h(V_c)$ is an alternative expression for f_c^h, which is more convenient for areas that require more complicated set theoretic expressions for their definitions, such as truncated Voronoi polygons.

This uniform density on Voronoi polygons, $\bar{f}_c^h$, is the vehicle we use in the present paper to 'translate' (density) mass associated with cells to geographically meaningful areas such as municipalities. [22]If municipality M is incident with Voronoi polygon V_c (that is, $M \cap V_c \neq \emptyset$), the mass that V_c 'donates' to M[23] is $\bar{f}_c^h\,|M \cap V_c|$. In this way all the Voronoi polygons incident with M donate to the mass of M, which is the sum of those 'donated' masses.

In case $\bar{c}$ is a terrestrial boundary cell we need to consider $V_{\bar{c}} \cap L$ instead of V_c. The mass associated with the truncated Voronoi polygon $V_{\bar{c}} \cap L$ is

$$m^h(V_{\bar{c}} \cap L) = \bar{f}_{\bar{c}}^h\,|V_{\bar{c}} \cap L|, \tag{22}$$

[21] In fact we are dealing with densities defined on cells, i.e. cell densities. But the density value at a cell is also viewed as a kind of mass. Hence the term 'density mass', to stress this.

[22] The implicit assumption is that all cells are omnidirectional. Due to lack of information of the type of cell (omnidirectional, unidirectional) this is a reasonable assumption. However, should more information be available about cells, this may result in a nonuniform distribution of cell density information on the corresponding Voronoi polygon.

[23] Or alternatively expressed: the mass that M inherits from V_c.

where the density $\bar{f}_{\bar{c}}^{h}$ is as defined in (21), for $c = \bar{c}$. The fact that the density is assumed to be uniformly distributed over a Voronoi polygon was used.

There are other instances in this paper where truncated Voronoi polygons are used, namely in case Voronoi densities are transformed into municipal densities. In that case, for a municipality M, we would find for a truncated interior cell c a similar expression as (22):

$$m^h(V_c \cap M) = \bar{f}_c^h \, |V_c \cap M|, \tag{23}$$

and for a boundary cell $\bar{c}$ intersecting with M:

$$m^h(V_{\bar{c}} \cap M \cap L) = \bar{f}_{\bar{c}}^h \, |V_{\bar{c}} \cap M \cap L|. \tag{24}$$

With this preparation we can move on to the transformation of geographic cell densities to Voronoi densities, which play a key role in the approach sketched in this paper.

6.4 Voronoi densities illuminated

In the present section some examples are shown of cells and the Voronoi partitioning they imply. In particular, the Voronoi polygons are of interest. It is of importance to distinguish those in the interior of the country from those at its border. The border polygons need special consideration as we have seen before.

We start looking at the cells, the generators of the Voronoi partition. In Figure 6.2 a map is shown with cells in The Netherlands.[24] It gives a good idea how dense this network is. Note that the cells are clustered in populous areas, i.e. where many people live, work, shop, go to school, go out, etc, as well as along main roads. The density is highest in the Randstad, a conurbation in the West of The Netherlands, containing major cities like Amsterdam, Rotterdam, The Hague and Utrecht. Other areas with high densities include Arnhem and Nijmegen, Breda, Eindhoven and the south of the Province Limburg. This latter area comprises of Maastricht (the western part) and the conurbation Parkstad Limburg (the eastern part).

In Figure 6.3 the cell density for the municipalities are presented as a colour map. In a sense this map is somewhat confusing as the size (area) of the municipalities is also of importance, not only the cell density. This makes direct comparison of Figure 6.2 and Figure 6.3 somewhat complicated. In particular in case of high densities and small areas or low densities and large areas.

[24] Information on cells in The Netherlands can be found at `https://www.antennebureau.nl/plaatsing-antennes/locaties-antennes-in-nederland` and at `https://www.gsmmasten.nl`.

Figure 6.2 Cells in the Netherlands, including the North Sea.

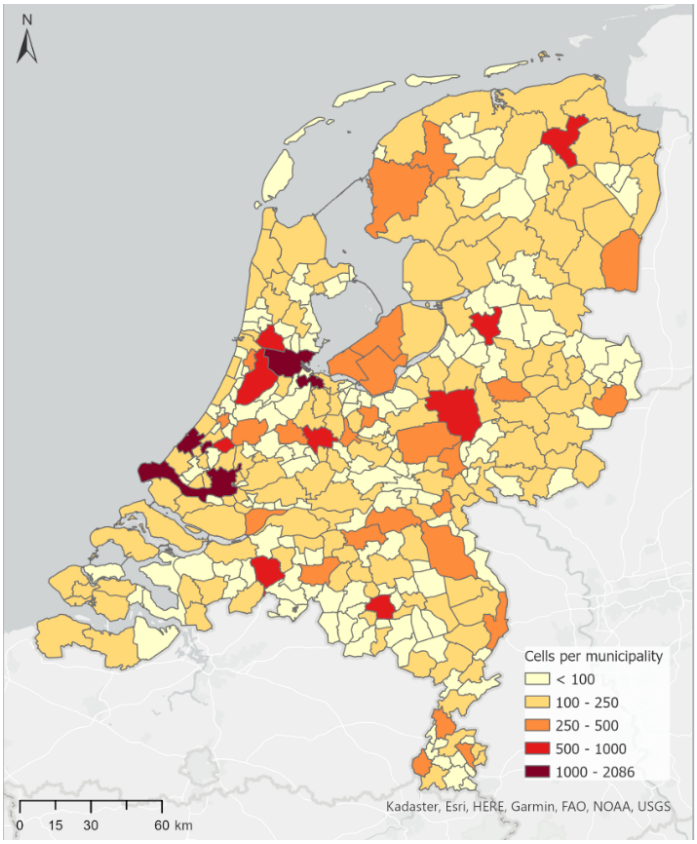

Figure 6.3 Heatmap of the number of cells per municipality in The Netherlands (in January 2023).

6.5 Boundary Voronoi polygons

In Section 6.1 we did not take into account that the Voronoi polygons we are actually interested in are not defined in the plane, but in a country L (viewed in a map), which restricts some of the polygons in the partitioning. But there is more to consider. This is best illustrated using a concrete example. So let L be (a suitable map) of The Netherlands. In particular, we are interested in its boundary. This can be divided into two parts:

1. **terrestrial**: borders with Belgium or with Germany.
2. **coastal**: bordering the North Sea.

The reason to distinguish between these parts is that in the first case there are likely cells available across the border, which, when used, transform the Voronoi polygons of border cells (on the Dutch side) to interior cells of 'normal' sizes. In the second case, there are offshore cells, of which there are relatively few and they are far apart. Using these offshore cells would not have the same effect of reducing the sizes of the Voronoi polygons concerned to 'normal' sizes. They would still be of a size that is too large. This would imply that the mass associated with these Voronoi polygons would be too big. See Section 6.5.3 for more discussion of this problem.

6.5.1 Characterizing boundary Voronoi polygons

In this section we want to concentrate on the difference between a Voronoi polygon of a boundary cell and such a polygon on an interior cell. The difference is of importance for our computations because in the latter case we can simply use the polygon itself in the calculations, whereas in the former case extra effort is required.

The following is a criterion that can be used to distinguish between both types of Voronoi polygons. Let V be a Voronoi in a partition generated by cells. Let L denote (a map of) the country (in our case, The Netherlands). If $V \cap L = V$ then V is an interior polygon. If $V \cap L \subset V$ then V is a boundary polygon.[25)]

6.5.2 Terrestrial boundary Voronoi polygons

In this case there are cells across the border in Belgium or Germany. They should be taken into account to make sure that the terrestrial border Voronoi polygons have 'normal' shapes and sizes, that is, comparable to interior Voronoi polygons. That these Voronoi polygons are partly covering Belgian or German soil is of no concern. In this case we make sure that the density computed is of better quality than in case only the part of the Voronoi polygons in The Netherlands (L) is taken into account: the area of these truncated Voronoi polygons would then be too small and hence also the mass to be associated with these geometric entities.

Figure 6.4 shows an area near the boundary of the Netherlands, in the south of the country, in the province of Limburg. It is a rather narrow strip of land bordering on Belgium as well as on Germany. Note the spiky Voronoi polygons corresponding to the boundary cells in that area. This phenomenon is due to the fact that no boundary cells in Belgium or Germany were used to 'tame' them by truncation. If such cells are used we obtain 'normal' looking Voronoi polygons, as Figure 6.5 shows. The boundary cells in The Netherlands are then also like interior cells. It should

[25)] It should be stressed that the symbol $\subset$ denotes a proper subset, one that is strictly smaller. So equality is excluded.

be stressed that the cells in Belgium or in Germany in Figure 6.5 are fictitious. They were used because no information on the location of cells in Belgium or Germany was at our disposal. These fictitious cells give an impression of their moderating influence on the shape of the spiky boundaries of Voronoi polygons associated with cells located in The Netherlands.

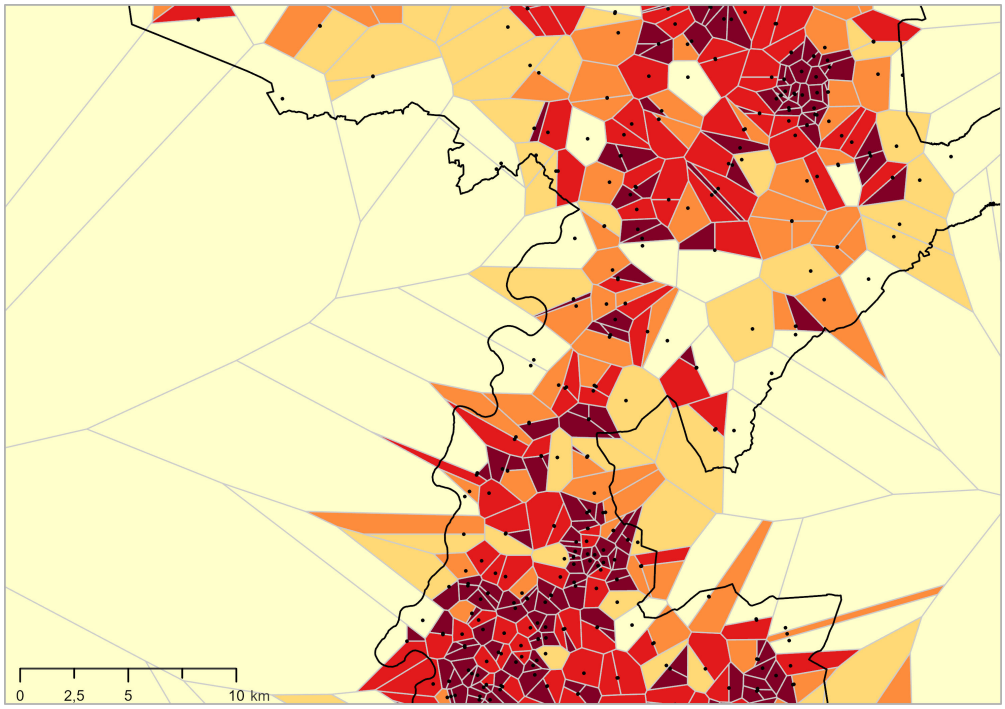

Figure 6.4 Voronoi polygons of terrestrial boundary cells near the borders of The Netherlands and its neighbours Belgium and Germany. No terrestrial boundary cells in these countries were used, to show the spikiness and oversized shape of some Voronoi polygons. The black line is the borderline of The Netherlands and these countries. A colour coded Voronoi density is shown as an example.

Remark The white areas in Belgium and Germany in Figure 6.5 should indicate that no communication information is available. Also no information on the location of cell phones in these countries needs to be known, except for the boundary cells in these countries near the Dutch border. These are only used to limit the size of the Voronoi boundary polygons at the Dutch side of the respective borders. But as no communication information was available about Belgian or German boundary cells, the corresponding Voronoi polygons were also left blank. The parts of these polygons that cover The Netherlands are therefore also left blank. On the other hand, if such information would have been available for these Belgian and German boundary cells it would be used for polygons covering parts of The Netherlands. So either way, there is an error. But as these concern relatively small pieces of territory, and also likely with little cell phone activity, their effects on the overall results are likely to be negligible. However, further investigation is needed to confirm (or reject) this assumption. □

6.5.3 Coastal boundary Voronoi polygons

For the offshore cells[26)] in the North sea near the coast (and removed from the Belgian and the German border) the situation is different. These cells are fewer and more widely spaced than

[26)] There are two types of cells near the coast: those on land (terrestrial) and those in the sea. The latter ones we call offshore cells, the former ones are called coastal (terrestrial) cells.

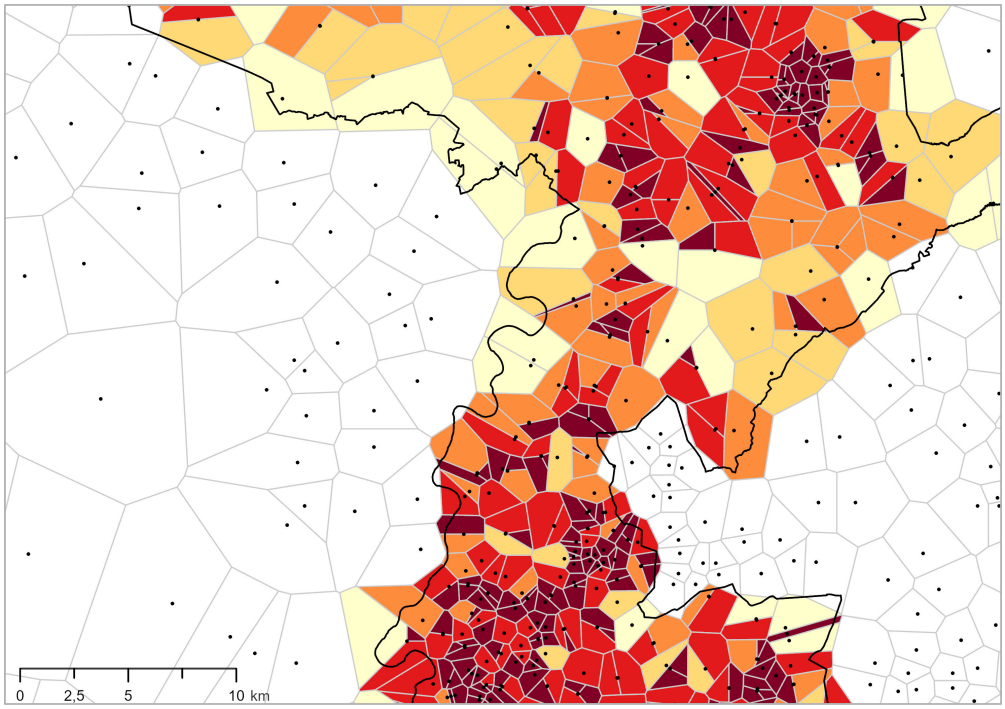

Figure 6.5 Voronoi polygons of terrestrial boundary cells in The Netherlands near the borders of Belgium and Germany with (fictitious) cells in these countries to show the truncating effect of the Voronoi polygons of boundary cells in The Netherlands. The black line indicates the border of The Netherlands and these neighbouring countries. A colour coded Voronoi density is shown as an example.

those on land (for obvious reasons). See Figure 6.6. They are used for communication with persons on board of ships (crews and passengers) in the area. These cells are probably unsuitable to create reasonably sized bounded polygons, as these Voronoi polygons tend to be rather big, and hence result in population densities for the coastal cells (on land) that are too low. Instead one could consider only the parts of the corresponding Voronoi polygons that intersect with land, as most active cell phones can be found there.

Another solution for the (terrestrial) coastal cells would be to assign the average area of the sizes of Voronoi polygons corresponding to interior cells, i.e. with Voronoi polygons bounded in size, possibly of cells situated not too far from the coast.

Figure 6.6 shows a coastal area in the north of The Netherlands, bordering the North Sea. Note that there are also cells in the North Sea. They are much less densely distributed than terrestrial cells, for obvious reasons. This explains the spikiness of the Voronoi polygons corresponding to (terrestrial) boundary cells. This time we should accept them, as they result from a real setting of cells.

Figure 6.6 also shows that we have spiky Voronoi polygons in the IJsselmeer, which used to be connected to the North Sea before it was closed off by the 'Afsluitdijk'.[27)]

In short, it is clear that the spiky boundary cells are a bit of an issue and should be considered more carefully. It is well possible that part of them are artifacts that can be left out of the picture

[27)] Literally translated 'Afsluitdijk' means 'closing dyke'. But actually the structure is a 'dam'. So 'Afsluitdijk' is, as a matter of fact, a misnomer.

as they are of little importance for the overall results. Or they are less harmful than they appear to be because they will be used truncated and only the parts that cover land will be used. But anyway, we shall not go into this issue here more deeply as we want to move on to other topics. This paper only wants to draw attention to them.

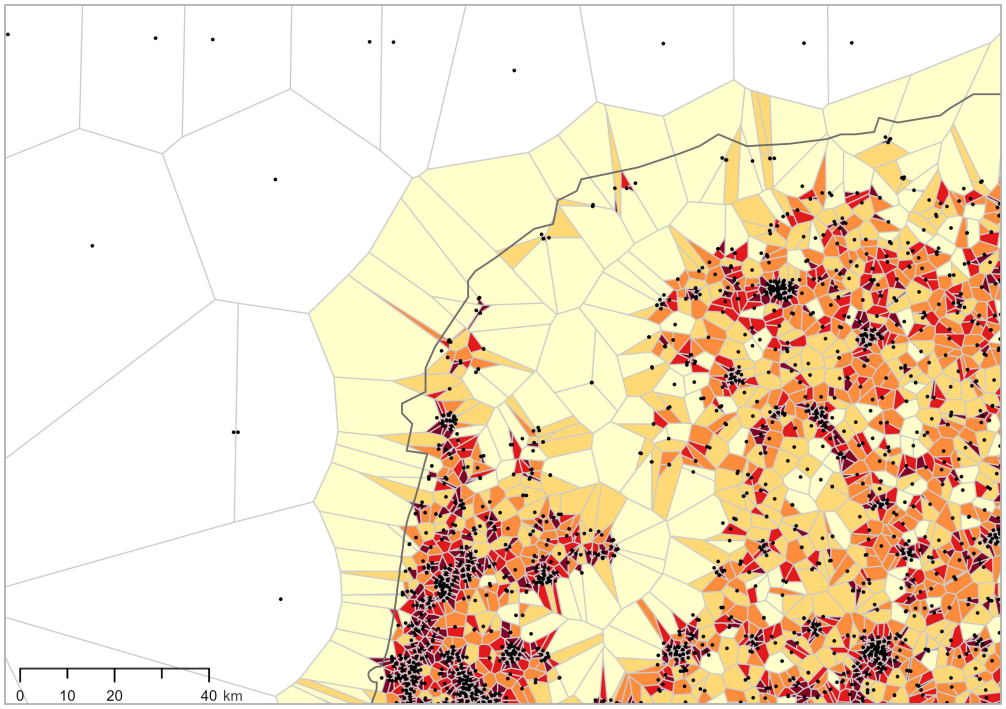

Figure 6.6 Coastal boundary Voronoi polygons near the North Sea. Cells located off shore (in the North Sea and the IJsselmeer) help to limit the size of the terrestrial boundary cells, but not as well as the cells across the border in Figure 6.5, as the density of the off shore cells is much less than that of the terrestrial cells in the area. A colour coded Voronoi density is shown as an example.

7 Municipal densities

The Voronoi partition generated by the locations of the cells as described in Section 6 is a means to an end. They are used to transform cell densities per hour block to similar densities defined on municipalities, or other meaningful geographic areas that partition the country. In fact, we only consider municipalities as such areas. For other meaningful geographic partitions the transformation of densities would be similar.

Two kinds of transformations are considered. The first transformation, in Section 7.2, is numerical in character, where a density value for each municipality is computed. As in case of a Voronoi density, the municipal density has a constant value for each municipality. These values may differ per municipality. This section describes the steps needed to compute the municipal density, for an hour block, from the Voronoi density for the same hour block. This process has to be repeated for each hour block. These municipal densities are considered separately. There is no attempt to understand how they evolved from each other. This aspect, concerning density flow, is considered somewhere else, namely in Section 14.

The second transformation, in Section 7.3, is graphical. Its goal is to produce a picture of the density. It obtained by smoothing a Voronoi density, using a graphical interpolation technique, and overlaying the result with the contours of the various municipalities, as a visual aid.

7.1 From Voronoi partition to municipal partition

We want to consider the transition of one partition of a set to another. As a concrete example, and one that is important for the present paper, we consider a Voronoi partition (generated by the locations of cells) and a partition in municipalities, both in country L (The Netherlands).

Let $\mathfrak{V} = \{V_1, \dots, V_n\}$ and $\mathfrak{M} = \{M_1, \dots, M_p\}$ be partitionings of L in Voronoi polygons and municipalities, respectively . Let

$$W_{ij} = V_i \cap M_j \tag{25}$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$. We are only interested in the nonempty sets W_{ij}. These sets are the building blocks for $\mathfrak{V}$ and $\mathfrak{M}$ in the sense that

$$\begin{aligned} V_i &= \bigcup_j W_{ij} \text{ , for } i = 1, \dots, n, \\ M_j &= \bigcup_i W_{ij} \text{ , for } j = 1, \dots, p, \end{aligned} \tag{26}$$

where the W_{ij} involved in the unions are all nonempty sets. These sets form the refinement of the partitions $\mathfrak{V}$ and $\mathfrak{M}$, which is denoted by $\mathfrak{V} \wedge \mathfrak{M}$. We have for $(i, j) \neq (k, l)$ that

$$W_{ij} \cap W_{kl} = \emptyset. \tag{27}$$

Let $r_{ij} = |W_{ij}|$, with W_{ij} given in (25). Let

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{np} \end{pmatrix}' = \begin{pmatrix} r_{11} & \cdots & r_{n1} \\ \vdots & \ddots & \vdots \\ r_{1p} & \cdots & r_{np} \end{pmatrix}. \tag{28}$$

R is used in Section 7.2, where it plays a part in the (linear) transformation from Voronoi densities to municipal densities.

7.2 From Voronoi density to municipal density

In the present section we show how the Voronoi density induced by cells can be transformed to a density on municipalities. It is an example of using the refinement of two partitions of the same set, in this case L. The partitions are $\mathfrak{V}$ of Voronoi polygons and $\mathfrak{M}$ of municipalities, as defined in Section 7.1. The Voronoi polygons are just auxiliary objects needed to obtain municipal densities from cell location densities, as municipalities are statistically meaningful geographic areas.

In analogy of equation (17) we can write

$$\begin{aligned} f_\mu^h &= \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{|W_{ij}|}{|V_i|}\mathbb{1}_{W_{ij}} f_{V_i}^h \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{|W_{ij}|}{|V_i|}\mathbb{1}_{W_{ij}} \frac{f_{c_i}^h}{|V_i|} \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{|W_{ij}|}{|V_i|^2}\mathbb{1}_{W_{ij}} f_{c_i}^h, \end{aligned} \tag{29}$$

where W_{ij} is defined in (25). We can write (29) in matrix form:

$$f_\mu^h = \left(\vec{1}_\mu\right)' \, RV^{-2} f_C^h, \tag{30}$$

where

$$\left(\vec{1}_\mu\right)' = \left(\mathbb{1}_{\mu_1}, \cdots, \mathbb{1}_{\mu_p}\right), \tag{31}$$

and where the μ_j, for $j = 1, \ldots, p$, represent the (geometric) municipalities, viewed as sets (polygons) situated in L. R is the matrix defined in (28), and V is the diagonal matrix:

$$V = \begin{pmatrix} |V_1| & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |V_n| \end{pmatrix}. \tag{32}$$

We can write (30) concisely as

$$f_\mu^h = \mathcal{M} f_C^h, \tag{33}$$

where

$$\mathcal{M} = \left(\vec{1}_{\mu}\right)' \, RV^{-2}. \tag{34}$$

$\mathcal{M}$ is only about the geometry/geography, namely the municipal partitioning of L. It only depends on the intersection of the Voronoi partition and the municipal partition. Of course, this independence of h only holds if the cell locations (and hence the Voronoi partitioning) and the definitions of the municipalities do not change during the observation period W. If they do, $\mathcal{M}$ changes accordingly. We shall assume that the Voronoi partitioning with the cell locations as generators and the municipal partitioning do not change during the observation period W.

The part of (33) consisting of f_C^h contains the statistical information. Now f_C^h is time dependent (hour block h) and is entirely unrelated to the municipal geometry.

Furthermore, $\mathcal{M} \geq 0$ and

$$\iota_p = \mathcal{M}\iota_n, \tag{35}$$

where $\mathcal{M}$ is of size $p \times n$, with p the number of municipailities and n the number of Voronoi polygons, which equals the number of cells. ι_p is the all ones column matrix of length p (as defined in (8)); likewise ι_n. $\mathcal{M}$ is a Markov like matrix, except that it is not square, as there are many more cells than municipalities.

In Figure 7.1 an example is shown of the transformation of a Voronoi density for an hour block to a municipal density for the same hour block, featuring the municipality Aalsmeer and the Voronoi polygons in the area. Those that intersect this municipality have been coloured with heavily accentuated colours.

In Figure 7.2 we see the same municipality as in Figure 7.1, namely Aalsmeer, with untruncated and truncated Voronoi polygons inside. For each truncated Voronoi polygon the associated weight is given, which is the ratio of its area and that of the original, untruncated polygon.

7.3 Heatmaps of Voronoi densities

The conversion in the present section is graphically oriented, whereas the one in Section 7.2 is numerical, intended for computations. There are several possibilities for such graphical conversions, intended to produce images that are smoothed, thus hiding abrupt changes in boundaries of areas and presented as colour-coded densities. Such colour-coded maps are called heatmaps.

If we want a density function that is smoother than a locally constant one (that is, constant on each Voronoi polygon) we need an extra processing step. There exist interpolation procedures that can be used for this purpose. There is nearest neighbour interpolation (see Appendix B.1) and natural neighbour interpolation (see Appendix B.2), which comes in (at least) two variants (by Sibson and by Laplace) that differ only in the weights used. In Chapter 6 of [8] the subject of smoothing is discussed in the context of Voronoi partitions.

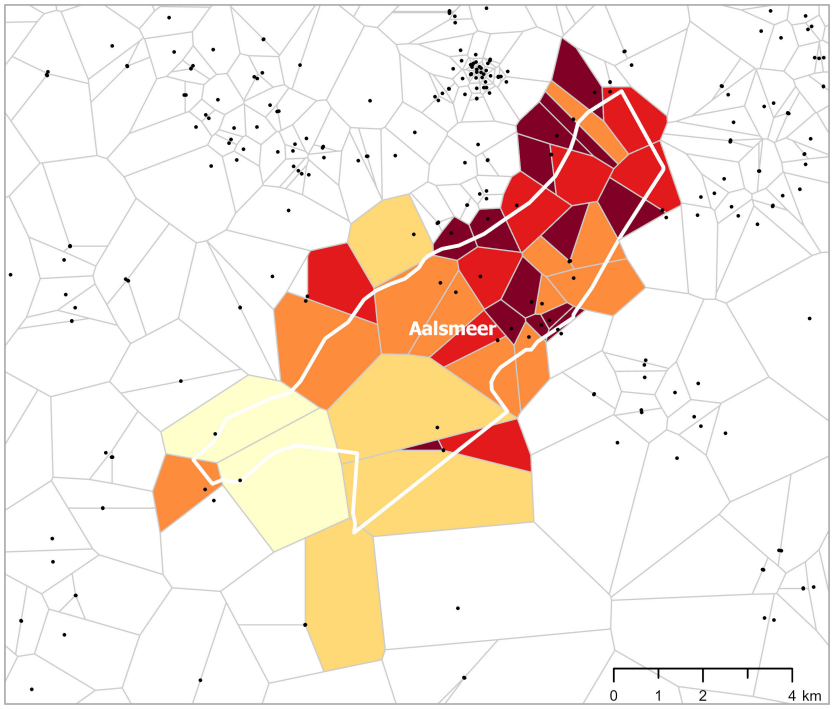

Figure 7.1 Municipality of Aalsmeer and the intersecting Voronoi polygons of neighbouring cells are shown. Nonintersecting Voronoi polygons have not been coloured.

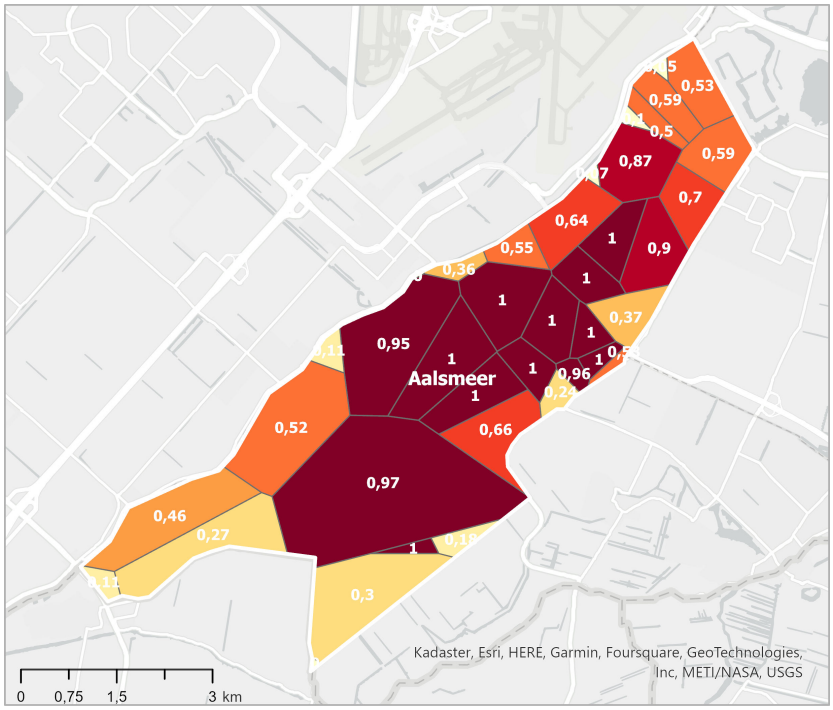

Figure 7.2 Truncated Voronoi polygons intersecting with the municipality of Aalsmeer are shown. The numbers inside the truncated polygons indicate the fraction of the area of the original Voronoi that this concerns (relative sizes). The colour coding is that of these relative sizes. The darker the colour, the larger the relative size.

In Figure 7.3 an example of a map with a colour coded Voronoi density is shown. The contour lines of municipalities have also been plotted, as a visual aid. In Figure 7.4 a smoothed version of this map is shown. Natural neighbour interpolation was used as a smoothing technique.

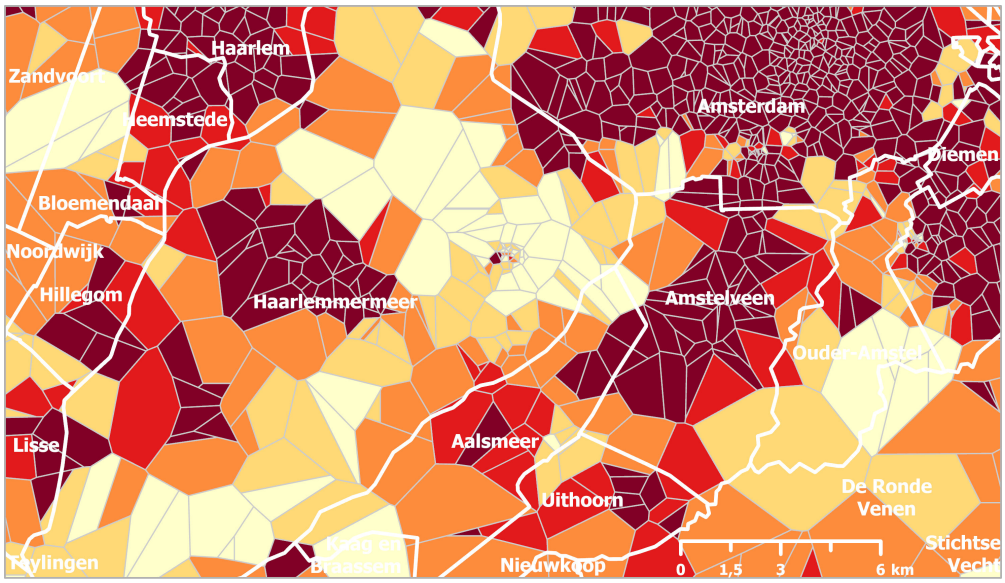

Figure 7.3 Voronoi polygons with the address density colour coded (the darker the colour, the higher the density value). Contours of municipalities have been added for visual support.

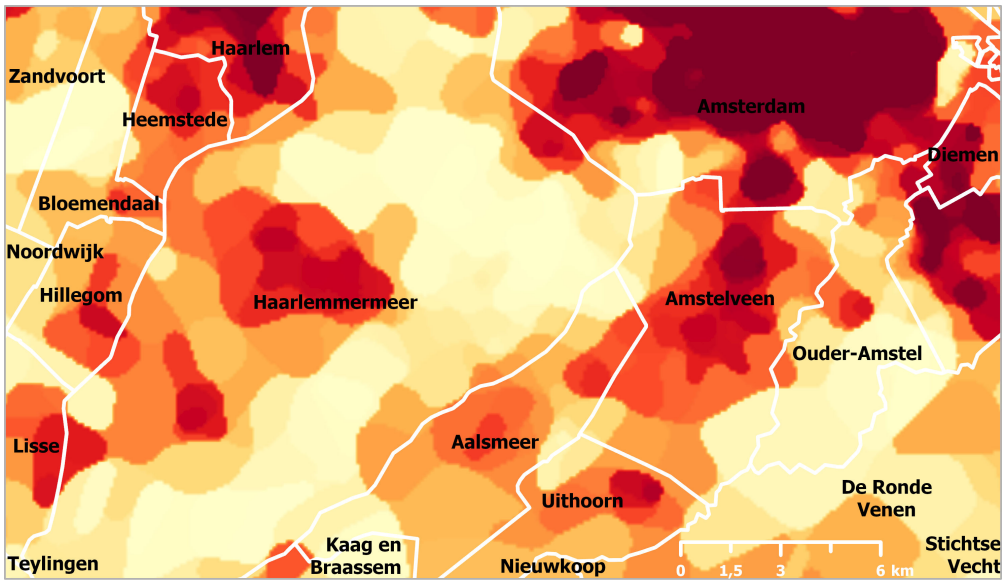

Figure 7.4 Heat map of a smoothed version of the map in Figure 7.3. The smoothing used natural neighbour interpolation. Contours of municipalities have been added for visual support.

In a more abstract sense, the smoothing of a Voronoi density f_V^h as defined in (17) into f_{SV}^h is a linear transformation $\mathcal{S}$ with

$$f_{SV}^h = \mathcal{S} f_V^h, \tag{36}$$

where S denotes a smoothing technique such as described in Appendix B, [28] such as natural neighbourhood interpolation. It should be noted that the linear transformation $\mathcal{S}$ does not depend on h.

It is not guaranteed that f_{SV}^h is a density, even though f_V^h is. It is certainly a nonnegative function. It can easily be made into a density by normalizing it.

8 Cell link digraph

The cell link digraph is a directed graph where the cells are the nodes. The arcs (c_1, c_2) indicate that cell c_2 can be reached from cell c_1 in at most 2 hours. The idea is that if a cell phone is active in c_1 in hour block h cell c_2 can be reached in hour block $h + 1$ if not even already in hour block h. The arcs are found on the basis of empirical data in the possession of the telecom provider. So the cell link digraphs gives a sense of proximity of cells in time units (hour blocks) that are relevant to the problem we study in the present paper.

In the present section we consider in some detail how the telecom provider could produce the cell links (arcs) from the source data. These data yield information about the reachability of cells within a limited time, in our case 2 subsequent hour blocks, say h and $h + 1$. So we should be looking for a cell in hour block h and which cells can be reached (by at least one cellphone) in hour block $h + 1$. In addition we should consider the possible links within one hour block, that is cells that are visited by cell phones within one hour block. Cells even closer to each other may then be revealed. As this is about reachability, the exact hour blocks h are not relevant. We start considering hour blocks h and hour block pairs h and $h + 1$, but then aggregate over these parameters.

To make things a bit more concrete we consider the example data in Appendix A. It is assumed that the production of the link data is done by the telecom provider, as the source data used should be considered sensitive as it is personal data. The link data released to the statistical office, however, are aggregate data obtained from these basic data: they are not about individual persons. There is also an aggregation over the various pairs of hour blocks h and $h + 1$ to obtain the link data, which also adds to the safety of the link data.

It should be stressed that the link data are derived from empirical data, that is, on the basis of observed movements of cell phones. Deriving such data from studying map information (including infrastructure) would in theory be possible, but would be much more complicated, more laborious to obtain and probably of poorer quality than follows from the empirical approach we take.

The link data are about reachability of cells, in a limited time. It has nothing to do with the intensity of links, only about possible reachability on the basis of actual observations. In Section 9 data concerning flow between cells is considered, that is also derived empirically. It provides

[28] Or another such technique, not mentioned here. There may also be nonlinear smoothing methods, but at this point we are not interested in these techniques, as they do not fit the linear transformation model that we are considering here.

each arc (cell link) with a number expressing the strength (intensity) of the flow between cells that are linked.

We first consider cells that were visited by at least one cell phone in different hour blocks. In a separate section we then consider the cells that are visited by cell phones in a single hour block.

8.1 Adjacent cells

In the present section we consider how cell phones travelled from a cell in hour block h to cells in hour block $h+1$. Each such combination defines an arc in the link digraph, which has the cells as its nodes. The arcs present direct transitions between cells. They indicate which cells can be reached from a given cell 'occupied' in one hour block and reached in the next hour block, that is its direct successor. As reachability[29] does not depend on the actual hour block pairs, the link data are obtained by aggregating the reachability results over all possible hour block pairs h and $h+1$.

For the moment we concentrate on a particular (c,h) combination and consider all the cell phone id's with that cell and hour block combination. Next, we consider all $(c,h+1)$ combinations. We are particularly interested in the cell phones appearing in the (c,h) combination that are also active in any cell d in hour block $h+1$. For then there is at least one active cell phone who passed from cell c to cell d from one hour block to the next. If there is such a cell phone we create an ordered pair (c,d) of cells. These ordered pairs will be arcs in the link digraph of cells.

Since we are looking for possible transitions from one active cell to the next, one hour block later, there are many more pairs of hour blocks to inspect, to be precise 23 as there are 24 hour blocks in a day (only the last hour block does not have a 'successor').

We now consider an example, namely the one represented in Appendix A. From Table A.1 we take two subtables with information that we need to derive the link data, which in fact is a digraph with the cells as nodes and where the arcs indicate direct transitions between cells for adjacent hour blocks. The results are presented in Table 8.1, which is about hour blocks h and $h+1$, and in Table 8.2, which is about hour blocks $h+1$ and $h+2$. The record numbers of the original table have been preserved for easy reference. They are not used in our computations.

From Table 8.1 we deduce Table 8.3, which shows only the pairs of cells for which (at least) one cell phone was active. These pairs of cells (in the order h, $h+1$) are arcs in the cell link digraph. Similarly we deduce Table 8.4 from Table 8.2.

Remark For simplicity we have chosen to use the same id's in Table 8.1 and Table 8.2. It would have been possible to use different and independent identifiers in these tables. But this would only yield extra safety within the telecom provider. These cell phone identities (original or surrogate) are not used in the data to be released to the statistical office. For our computations such deliberations are not needed. The identifiers in each table are only used to collect the data that belong to the same cell phone. The actual identity of a cell phone is of no importance. Surrogate keys could be used instead. □

[29] As a potentiality, a possibility. We do not consider the situation whereby reachability is time dependent. This happens in reality, but will be left out of our considerations for simplicity's sake.

rec	h bl	cell	id	rec	h bl	cell	id
1	h	c_1	id_1	13	h+1	c_1	id_2
2	h	c_2	id_3	14	h+1	c_2	id_2
3	h	c_2	id_5	15	h+1	c_3	id_3
4	h	c_3	id_5	16	h+1	c_3	id_5
5	h	c_3	id_6	17	h+1	c_4	id_5
6	h	c_4	id_6	18	h+1	c_4	id_6
7	h	c_4	id_8	19	h+1	c_4	id_8
8	h	c_5	id_8	20	h+1	c_6	id_8
9	h	c_6	id_8	21	h+1	c_6	id_{11}
10	h	c_6	id_{11}	22	h+1	c_6	id_{12}
11	h	c_6	id_{12}	23	h+1	c_7	id_{12}
12	h	c_7	id_{12}				

Table 8.1 Records (rec) of active cell phones (id) at cells (cell) in hour blocks (h bl) h and $h+1$.

rec	h bl	cell	id	rec no	h bl	cell	id
13	h+1	c_1	id_2	27	h+2	c_4	id_5
14	h+1	c_2	id_2	28	h+2	c_5	id_6
15	h+1	c_3	id_3	29	h+2	c_5	id_7
16	h+1	c_3	id_5	30	h+2	c_7	id_8
17	h+1	c_4	id_5	31	h+2	c_7	id_9
18	h+1	c_4	id_6	32	h+2	c_1	id_9
19	h+1	c_4	id_8	33	h+2	c_1	id_{10}
20	h+1	c_6	id_8	34	h+2	c_2	id_{11}
21	h+1	c_6	id_{11}	35	h+2	c_3	id_{11}
22	h+1	c_6	id_{12}	36	h+2	c_4	id_{11}
23	h+1	c_7	id_{12}	37	h+2	c_5	id_{12}
24	h+2	c_2	id_1	38	h+2	c_6	id_{12}
25	h+2	c_2	id_2	39	h+2	c_7	id_{12}
26	h+2	c_3	id_4				

Table 8.2 Records (rec) of active cell phones (id) at cells (cell) in hour blocks (h bl) $h+1$ and $h+2$.

From Table 8.3 we deduce the following arcs of the cell link digraph, including loops:

(c_2, c_3), (c_2, c_4), (c_3, c_3), (c_3, c_4), (c_4, c_4), (c_4, c_6), (c_5, c_4), (c_5, c_6), (c_6, c_4), (c_6, c_6), (c_6, c_7), (c_7, c_6), (c_7, c_7).

The first cell of each pair pertains to hour block h, whereas the second one pertains to hour block $h+1$.

From Table 8.4 we deduce in the same way as before the following arcs of the cell link digraph, including loops:

(c_1, c_2), (c_2, c_2), (c_3, c_4), (c_4, c_4), (c_4, c_5), (c_4, c_7), (c_6, c_2), (c_6, c_3), (c_6, c_4), (c_6, c_5), (c_6, c_6), (c_6, c_7), (c_7, c_5), (c_7, c_6), (c_7, c_7).

Combining these two sets of arcs yields, after deduplication:

(c_1, c_2), (c_2, c_2), (c_2, c_3), (c_2, c_4), (c_3, c_3), (c_3, c_4), (c_4, c_4), (c_4, c_5), (c_4, c_6), (c_4, c_7), (c_5, c_4), (c_5, c_6), (c_6, c_2), (c_6, c_3), (c_6, c_4), (c_6, c_5), (c_6, c_6), (c_6, c_7), (c_7, c_5), (c_7, c_6), (c_7, c_7).

id	h	$h+1$
id_3	c_2	c_3
id_5	c_2, c_3	c_3, c_4
id_6	c_3, c_4	c_4
id_8	c_4, c_5, c_6	c_4, c_6
id_{11}	c_6	c_6
id_{12}	c_6, c_7	c_6, c_7

Table 8.3 Cell phones (id) and the cells (c_j) active in hour blocks h and $h+1$.

id	$h+1$	$h+2$
id_2	c_1, c_2	c_2
id_5	c_3, c_4	c_4
id_6	c_4	c_5
id_8	c_4, c_6	c_7
id_{11}	c_6	c_2, c_3, c_4
id_{12}	c_6, c_7	c_5, c_6, c_7

Table 8.4 Cell phones (id) and the cells (c_j) active in hour blocks $h+1$ and $h+2$.

If there were more pairs of adjacent hour blocks, they would also yield arcs, which should be added to this list, if not already present. In Section 8.2 we discuss that there are even more arcs of the cell link digraph that can be retrieved from Tables 8.3 and 8.4.

Remark A move from cell c to itself may be achieved by not moving at all, provided this is physically possible.[30] But it may also be achieved, for example, by an active cell phone communicating via c which then moves, de-activated, to any other place, and then is back near c in hour block $h+1$ when it is activated. This, of course, holds for any cell c. Note that this fact follows from a reasoning using common knowledge, rather than from empirical observation. Such a movement cannot be detected by the method we are using. □

In the present section we found arcs by considering cells in which a cell phone was present in adjacent hour blocks. But this is not all: within a single hour block it is possible for a cell phone to be active at different cells. Section 8.2 considers this possibility.

8.2 Nearby adjacent cells

In Section 8.1 we considered links (c_i, c_j) in the cell link digraph with c_i from an hour block k and c_j from the next hour block $k+1$. But in addition to these links we also have those where c_i and c_j are from the same hour block, as they also can be reached within two hours. If we inspect Table 8.3 we derive the following such arcs:

(c_2, c_3), (c_3, c_2), (c_3, c_4), (c_4, c_3), (c_4, c_5), (c_4, c_6), (c_5, c_4), (c_5, c_6), (c_6, c_4), (c_6, c_5), (c_6, c_7), (c_7, c_6).

Note that for each pair $\{a, b\}$ of such cells we have assumed they they imply the arc (a, b) as well as the arc (b, a). This actually is based on an assumption: we know at least one of these is an arc, but we do not know which one. One could delve deeper into the data to find out, which one is an

[30] An exception to this would be a cell phone in an area at which one cannot stay put and can only pass through. For instance in case of a cell next to a train track. Such cells are probably rare (in The Netherlands), if they exist at all.

arc (and possibly both) but we think the present method is acceptable and has the advantage that it needs no extra work. Likewise we can deduce from Table 8.4 the following such arcs:

(c_1, c_2), (c_2, c_1), (c_2, c_3), (c_2, c_4), (c_3, c_2), (c_3, c_4), (c_4, c_2), (c_4, c_3), (c_4, c_6), (c_5, c_6), (c_5, c_7), (c_6, c_4), (c_6, c_5), (c_6, c_7), (c_7, c_5), (c_7, c_6).

If we combine both sets of newly found arcs with the one obtained in Section 8.1 we find, after deduplication:

(c_1, c_2), (c_2, c_1), (c_2, c_2), (c_2, c_3), (c_3, c_3), (c_2, c_4), (c_3, c_2), (c_3, c_4), (c_4, c_3), (c_4, c_4), (c_4, c_5), (c_4, c_6), (c_4, c_7), (c_5, c_4), (c_5, c_6), (c_5, c_7), (c_6, c_2), (c_6, c_3), (c_6, c_4), (c_6, c_5), (c_6, c_6), (c_6, c_7), (c_7, c_5), (c_7, c_6) (c_7, c_7).

We can represent this in a more concise form as an adjacency matrix. See Table 8.5.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	1	0	0	0	0	0	0	0
c_2	1	1	1	1	0	0	0	0	0
c_3	0	1	1	1	0	0	0	0	0
c_4	0	0	1	1	1	1	1	0	0
c_5	0	0	0	1	0	1	1	0	0
c_6	0	1	1	1	1	1	1	0	0
c_7	0	0	0	0	1	1	1	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 8.5 Adjacency matrix of the cell link digraph.

It should be pointed out that the adjacency matrix in Table 8.5 concerns a toy example, so it has some odd features. For instance, the fact that some rows and colums contain only 0's implying that certain cells are not linked to any other cell (such as cells c_8 and c_9). For an adjacency matrix based on real data this is highly unlikely. However, the asymmetry of the adjacency matrix in the example is likely to hold for a real cell network as well. This is also the case for the existence of loops, corresponding to transitions to the same cell. Having loops is not a property of flow networks. Instead, such networks have the property that for each node in the network holds that total inflow equals total outflow.[31)] This property precludes that nodes have a capacity to store 'mass'. As the cell networks we consider in this paper do have the 'mass' storing capacity, we are dealing with somewhat different flow networks here.

9 Cell density flow

The cell density flow is the basic density flow in the present paper from which all other density flows in this paper are derived: the geometric cell density flow, the Voronoi flow and the municipality flow. The cell density flow is derived from the intermediate data provided by the telecom provider to the statistical office (see Section 3.3). From this information Markov matrices can be derived, indicating the flow from two consecutive hour blocks h and $h + 1$, for $h = 1, \ldots, 23$.

31) Which is called Kirchhoff's law in the theory of electrical networks.

9.1 Deriving the Markov matrices

In this section we describe data on the flow between cells, i.c. Markov matrices for each pair of consecutive hour blocks h and $h+1$, for $h = 1, \ldots, 23$, assuming the time window is a full day. To obtain these data the method to obtain the link data as considered in Section 8 should be refined. We consider the construction of Markov matrices by using the example data of Appendix A. It should be stressed that the telecom provider should prepare these data and deliver them to the statistical office for use. In this way the safety of the data is guaranteed. The Markov matrices can be considered safe. The intermediate data is not safe, but this is not a problem as the telecom provider is the only actor to handle them, who is also the owner of the data from which they are produced.

We now go back to the two sets of arcs in the cell link digraph, one derived from Table 8.3 and the other from Table 8.4. But we now look differently at them: we are not interested in links between cells but in flows between cells occurring between adjacent hour blocks, in our case h and $h+1$ as well as $h+1$ and $h+2$. The information we need can be derived from Table 8.3 for hour blocks h and $h+1$ and from Table 8.4 for hour blocks $h+1$ and $h+2$. This time we cannot pool the data, as flow is a time dependent quantity. This time we do not deduplicate cells as in the cell link case, but, quite to the contrary, we count the numbers of duplications. The results can be found in Tables 9.1 and 9.2.[32)]

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	0	0	0	0	0	0	0	0
c_2	0	0	2	1	0	0	0	0	0
c_3	0	0	1	2	0	0	0	0	0
c_4	0	0	0	2	0	1	0	0	0
c_5	0	0	0	1	0	1	0	0	0
c_6	0	0	0	1	0	3	1	0	0
c_7	0	0	0	0	0	1	1	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 9.1 Flow matrix for the hour blocks h and $h+1$.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	1	0	0	0	0	0	0	0
c_2	0	1	0	0	0	0	0	0	0
c_3	0	0	0	1	0	0	0	0	0
c_4	0	0	0	1	1	0	1	0	0
c_5	0	0	0	0	0	0	0	0	0
c_6	0	1	1	1	1	1	2	0	0
c_7	0	0	0	0	1	1	1	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 9.2 Flow matrix for the hour blocks $h+1$ and $h+2$.

From the flow matrices in Tables 9.1 and 9.2 the Markov matrices in Tables 9.3 and 9.4 can be obtained. Without further information these matrices describe the probability p_{cd} that a cell phone at a certain cell c in hour block h (respectively, $h+1$) moves to a cell d in hour block $h+1$ (respectively, $h+2$). This creates the picture of the flow of population density following a time

[32)] There is little duplication, however, in this toy example!

dependent Markov chain, as the Markov matrices associated with adjacent hour blocks are not constant but are allowed to change over time. The lack of memory of the diffusion process is a result of the fragmented view of the flow that has been created on purpose: only from an hour block h to its immediate successor $h+1$ for $h = 1, \ldots, 23$, as no cell phone is followed over time. This is precisely the view underlying a Markov process. It is time dependent because the various Markov matrices involved can vary over time.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	0	0	0	0	0	0	0	0
c_2	0	0	$2/3$	$1/3$	0	0	0	0	0
c_3	0	0	$1/3$	$2/3$	0	0	0	0	0
c_4	0	0	0	$2/3$	0	$1/3$	0	0	0
c_5	0	0	0	$1/2$	0	$1/2$	0	0	0
c_6	0	0	0	$1/5$	0	$3/5$	$1/5$	0	0
c_7	0	0	0	0	0	$1/2$	$1/2$	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 9.3 Markov matrix for the hour blocks h and $h+1$.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	1	0	0	0	0	0	0	0
c_2	0	1	0	0	0	0	0	0	0
c_3	0	0	0	1	0	0	0	0	0
c_4	0	0	0	$1/3$	$1/3$	0	$1/3$	0	0
c_5	0	0	0	0	0	0	0	0	0
c_6	0	$1/7$	$1/7$	$1/7$	$1/7$	$1/7$	$2/7$	0	0
c_7	0	0	0	0	$1/3$	$1/3$	$1/3$	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 9.4 Markov matrix for the hour blocks h and $h+1$.

That the diffusion proces we are dealing with is not stationary is clear if we realize that the movements during a working day in the morning are different from those in the evening. Many workers travel to their work and in the evening many of them travel back. During the wekend one may see different patterns: people tend to go shopping on Saturday and some time later they return home. On Sunday they may visit a museum, a sporting event, a church, etc. and then return home. Of course, the weather has an influence on when, how and where people travel. Tourists and visitors to the country have their own behavioural patterns. All this buzzing around is reflected in the Markov matrices.

9.2 Transformations between consecutive cell densities

Here we consider the link between cell density f_C^h and f_C^{h+1}. If M_C^h denotes the Markov matrix that describes the transition between the cell densities at hour blocks h and $h+1$, we can write

$$f_C^{h+1} = \left(M_C^h\right)' f_C^h. \tag{37}$$

It should be stressed that equations (37), for $h = 1, \ldots, 23$, in practice only hold approximately, because the f_C^h and M_C^h are independent estimates which are unlikely to fit seamlessly, even in

case measurement errors are absent. In Appendix E some suggestions are provided concerning the modification of f_C^h and M_C^h for $h = 1, \dots, 23$, in such a way that the equations (37) are satisfied exactly.

9.3 Inactive cell phones and missing flow information

As remarked in Section 4.2 we should not ignore the fact that the population of cell phones is an open one and that a cell phone that is not active has died, so to speak, but its whereabouts are not known. It may have left (or entered) the country but it may also have been switched off (or on). In case it was switched off it still has an (unknown) location in the country (i.e. a cell, or even several cells, if it is moving) in which it would be localized if switched on. So it is missing information problem that we are dealing with. In case of cell flow information we may see the same problem.

This observation is closely linked to the missing data problem raised in Section 2.4.2. In fact, we are dealing with two phenomena which both lead to missing data, although they are basically different. On the one hand we have cell phones that are switched on or off, but that remain in the country (or the area covered by the cells).[33] On the other hand we have cell phones that enter or leave the cell network. In the first case we could say that the cell phones are still present but their positions are unknown. In the second case it is more appropriate to talk about an open population, as cell phones physically enter or leave the area covered by the cells in the network (roughly the country itself).[34]

The problem is that the two situations cannot be distinguished if it is possible that a cell phone is present in the country L in hour block h and is outside L in hour block $h + 1$. If this would be the case with a cell phone we would not know if it has left the country or has been switched off. In case the country is The Netherlands it seems possible to exit the country in the situation sketched from pretty every location in the country in the given time span. Similar things can be said about entering the country. If half hour blocks would be used, or even shorter time periods, there would be areas that are too far removed from the border to exit or enter the county in a short time span. In a bigger country or when shorter time blocks are used, there are places from which one can (under normal circumstances) not leave or enter the country. Then one knows (almost for certain) that cell phones starting or arriving there one time block later, that the cell phone must have been switched off. Likewise it is reasonable to assume that cell phones disappearing or appearing in the next time block have probably left the country or entered it. One could then proceed with modelling both phenomena and estimate parameters from the data. Using observed travelling patterns one could use this to make estimates for the missing flow information. However, we shall not elaborate this suggestion, as in case of The Netherlands it cannot be applied fruitfully: the country is too small when the time blocks used are in fact hour blocks, as we have assumed from the outset. We only give some initial ideas for the approach.

So suppose we distinguish between the two causes for (dis)appearing cell phones:

– by being switched on or off by their users.

[33] For simplicity we assume that the area covered by the cell network and the area covered by the country L coincide. But this is only approximately true.

[34] That cell phones outside the country also have locations is of course true but irrelevant, as our scope is only the cell phones within the country L.

– through entering or leaving the cell network, which we assume roughly to correspond with entering or leaving the country.

The first phenomenon can be studied in 'interior cells', cells for which the second phenomenon cannot occur because one is too far removed from the boundary (more than a time block's travel by the fastest mode of transport that would locally be available, say a train or a car).

Another possibility is to treat the unknown start (previous cell) of a cell phone (when switched on) or the unknown goal (next cell) as missing (cell) values and impute these. A simple model would consider the switching on or off of a cell phone as a random event, which happens independently of the start or goal cell. This would result in a missing data model in which the start cell or the goal cell equals the same pattern as in case of fully observed cells (provided these are available). See Appendix D for some ideas on such an imputation model. For a general discussion of missing data and how to deal with them in statistical data see [6].

This imputation would, however, not necessarily result in 'population mass' preservation. A switched off cell phone may be switched on for the first time not immediately, but several time blocks later. It is not the intention to follow a cell phone with an imputed cell position for more than one step, as this would be highly speculative, as the data for this are not available. This imples that you consider a path generated by a Markov chain, generating the next direction by drawing from the appropriate set of transition probabilities. Of course, this is not how individuals travel in reality. But it is a way to describe the movements of a collective of individuals.

Remark A special case is at night when many persons are not active on their cell phones because they are asleep. So many missing data for cell phone locations is an inherent weakness of the method considered in this paper. Using only the relatively few active cell phones is possibly risky. To correct for this some extra modelling work is needed, with additional input to be provided by the telecom provider, such as the last observed cell locations of cell phones, in aggregate form. We shall not study this problem in the present paper, but reserve it for future research. □

10 Geometric cell density flow

The geometric cell density flow is strongly linked to the cell density flow. In this case the new aspect is that the flow has obtained a geometric dimension. The development at this level is in itself not important. It concerns an intermediate step, which however is important. It allows two things to be realized: computation of the Voronoi density flow (see Section 11) and computation of the municipality density flow (see Section 13). In the former case, the locations of the cells can be used as generators of a Voronoi partition. In the latter case, the locations of the cells are used to determine the municipality each cell is located in.

Using (14) we have for the link between the geometric cell densities of hour blocks h and $h+1$:

$$f_{GC}^{h+1} = \left(\vec{\mathbb{1}}_{GC}\right)' f_C^{h+1} = \left(\vec{\mathbb{1}}_{GC}\right)' \left(M_C^h\right)' f_C^h = \left(M_C^h \vec{\mathbb{1}}_{GC}\right)' f_C^h. \tag{38}$$

where (37) was used to link f_C^{h+1} and f_C^h. Note that (38) consists of a part (namely $\vec{\mathbb{1}}_{GC}$) that is about the location of the cells and a part (namely $\left(M_C^h\right)' f_C^h$) that is about the cell density flow. We see a similar split between geometry and cell density flow in case of Voronoi density flow in Section 11 and the municipality flow in Section 13.

If we put

$$\mathcal{M}_{GC}^h = \left(M_C^h \vec{\mathbb{1}}_{GC}\right)'. \tag{39}$$

then we can write (38) concisely as

$$f_{GC}^{h+1} = \mathcal{M}_{GC}^h f_C^h. \tag{40}$$

We could explore the geographic cell flow a bit more, but it does not seem worth the effort. The geographic cell density is not very interesting by itself. It is just an intermediate density to derive Voronoi density and from this the municipal density. The corresponding flows are important and we will consider these in Sections 11 and 13.

11 Voronoi density flow

Voronoi density flow is very closely linked to cell density flow as there is a 1-1 relationship between cells and cell locations and between cell locations and Voronoi polygons. Using (20) we have for hour block $h+1$

$$f_V^{h+1} = \mathcal{V}\left(M_C^h\right)' f_C^h. \tag{41}$$

In (41) there is also a neat separation of information about the geometry of the Voronoi partition (namely $\mathcal{V}$) and information about the cell density (namely $\left(M_C^h\right)' f_C^h$), more particularly, about its dynamics. The geometric structure, i.e. the Voronoi partitioning, is assumed to be static in the present paper.

By combining (41) and (20) we derive the following difference equation:

$$f_V^{h+1} - f_V^h = \mathcal{V}\left(\left(M_C^h\right)' - I_n\right) f_C^h, \tag{42}$$

where I_n is the $n \times n$ identity matrix.

From (41) also follows:

$$f_V^{h+1} = \mathcal{V}\left(\prod_{t=1}^{h}\left(M_C^{h-t+1}\right)'\right) f_C^1 = \mathcal{V}\left(\prod_{t=1}^{h} M_C^t\right)' f_C^1 \tag{43}$$

If we put

$$\mathcal{W}_V^h = \mathcal{V}\left(M_C^h\right)'. \tag{44}$$

then we can write (41) as

$$f_V^{h+1} = \mathcal{W}_V^h f_C^h. \tag{45}$$

$\mathcal{W}_V^h$ is a linear transformation, depending on both geometric/geographical and statistical information.

12 Municipal link digraph

In Section 8 we considered the cell link digraph. In the present section we want to use this digraph to derive a municipal link digraph. We start with the adjacency matrix of the cell link digraph. The idea is now to partition the (terrestrial) cells into groups according to the municipality in which each of them is located. We first present a small example illustrating the idea.

Example We consider the adjacency matrix in Table 8.5 for the toy cell link digraph in Section 8. In Table 12.1 we have indicated the cells grouped into the clusters $\mathcal{C}_1 = \{c_1, c_2, c_3\}$, $\mathcal{C}_2 = \{c_4, c_5\}$, $\mathcal{C}_3 = \{c_6, c_7\}$ and $\mathcal{C}_4 = \{c_8, c_9\}$.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_1	0	1	0	0	0	0	0	0	0
c_2	1	1	1	1	0	0	0	0	0
c_3	0	1	1	1	0	0	0	0	0
c_4	0	0	1	1	1	1	1	0	0
c_5	0	0	0	1	0	1	1	0	0
c_6	0	1	1	1	1	1	1	0	0
c_7	0	0	0	0	1	1	1	0	0
c_8	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	0	0	0	0	0

Table 12.1 Partitioned adjacency matrix of the cell link digraph defined in Section 8, with respect to the clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ and $\mathcal{C}_4$.

In Table 12.2 the adjacency matrix after clustering the cells in Table 12.1 is shown. In this case it is not very interesting as it is very small. And it has only binary information as a link exists or it does not.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$
$\mathcal{C}_1$	1	1	0	0
$\mathcal{C}_2$	1	1	1	0
$\mathcal{C}_3$	1	1	1	0
$\mathcal{C}_4$	0	0	0	0

Table 12.2 Adjacency matrix for the clusters of cells $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ derived from Table 12.1.

In Table 12.3 the strength matrix is shown that is derived from Table 12.1. In this matrix the strength of each link is expressed by the number of cells that are contained in each part of this latter table. This gives an impression about the intensity of links.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$
$\mathcal{C}_1$	6	2	0	0
$\mathcal{C}_2$	1	3	4	0
$\mathcal{C}_3$	2	3	4	0
$\mathcal{C}_4$	0	0	0	0

Table 12.3 Strength matrix for the clusters of cells $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ derived from Table 12.1.

The strength matrix in Table 12.3 can also be expressed in a relative fashion, in such a way that all row totals add up to 1. This yields Table 12.4 instead of Table 12.3.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$
$\mathcal{C}_1$	$3/4$	$1/4$	0	0
$\mathcal{C}_2$	$1/8$	$3/8$	$1/2$	0
$\mathcal{C}_3$	$2/9$	$1/3$	$4/9$	0
$\mathcal{C}_4$	0	0	0	0

Table 12.4 Relative strength matrix for the clusters of cells $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ derived from Table 12.1.

□

With this example in mind, it is straightforward to define the municipal link graph. This time we start wil the cell link digraph. For each municipality we cluster the cells inside it. Then for each cluster we determine if it is empty or not, coded as 0 or 1, respectively. This yields the municipal link digraph we are looking for.

13 Municipal density flow

As in case of cell density flow, we can consider municipal density flow. In both cases the flow is described by appropriate Markov matrices. We first consider the formal aspects of the municipal flows, assuming Markov matrices to describe the flow for each pair of consecutive hour blocks. This is in complete analogy of the cell flow density case, and mainly serves the purpose of establishing the necessary concepts and notation. Then, in a separate section, we consider the link between the Markov matrices at the cell level and the corresponding Markov matrices at the municipal level.

13.1 Linking municipal densities of consecutive hour blocks

We now consider f_μ^h and f_μ^{h+1}, the municipal densities for hour blocks h and $h+1$, respectively. The transformation linking f_μ^h to f_μ^{h+1} is the Markov matrix M_μ^h. So we have:

$$f_\mu^{h+1} = \left(M_\mu^h\right)' f_\mu^h. \tag{46}$$

M_μ^h is an $m \times m$ Markov matrix, where m denotes the number of municipalities. So $M_\mu^h \geq 0$ and $M_\mu^h \, \iota_m = \iota_m$, the all 1's column vector of length m. This matrix M_μ^h depends on h as the dispersal of cell phones is likely to follow the traffic pattern, for all modes of transport available combined. In the morning a lot of people go to work and in the evening they return home, showing the reversal dispersion pattern as in the morning. But this is a pattern only for those with a day job, at a time when there are no holidays. People who work nightshifts, or who work from home, or who are without a job, or tourists, they all have different travel patterns during the day. Of course, these groups are not homogeneous and people tend to have individual travel patterns. A combination of all these patterns is reflected in these matrices M_μ^h for the various hour blocks h.

The Markov matrices M_μ^h are not independent from their counterparts M_C^h in the cell case. In Section 13.2 we consider this link more closely.

Remark It should be noted that (46) holds for population values for the densities f_μ^h, f_μ^{h+1} and the Markov matrices M_μ^h. But in practice these are not known, only estimates thereof. For these estimates it is not guaranteed that (46) holds. This requires some adjustments. There are various possibilities: one can take the Markov matrices obtained from the observed data and consider them as good approximations of the real versions. Then one needs to adjust the densities f_μ^h. Or one considers both the densities and the Markov matrices as not accurate above, and then sets out to adjust them both. Or one considers the densities f_μ^h as sufficiently precise and then tries to adjust the Markov matrices. A priori, the last option seems to be the least likely. The present paper is not able to deal with this issue, as no data are available. It is therefore also not clear to which extent this is a problem in practice, that is, how much (46) is violated and how much correction is needed. □

13.2 Linking Markov matrices for cell and municipal density flows

In this section we consider the transformation of a cell density flow, as discussed in Section 9, to a municipal cell flow. This can be achieved by aggregating Markov matrices describing cell flow density in an appropriate manner to be shown. It can be likened to the case of deriving municipal links from cell links in Section 9. We start illlustrating the approach with an example.

Example Suppose we have a 5×5 Markov matrix M as presented in (47), with states $1, \ldots, 5$. By grouping states $1, 2, 3$ into state A and states $4, 5$ into state B, M induces a Markov matrix $\bar{M}$, as given in (50).

$$M = \left(\begin{array}{ccc|cc} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ \hline p_{41} & p_{42} & p_{43} & p_{44} & p_{45} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} \end{array}\right) = \left(\begin{array}{c|c} M_{11} & M_{12} \\ \hline M_{21} & M_{22} \end{array}\right), \tag{47}$$

where M_{11}, M_{12}, M_{21} and M_{22} are submatrices of M of order 3×3, 3×2, 2×3 and 2×2, respectively. We have

$$M\iota_5 = \iota_5, \tag{48}$$

and hence

$$\left(\begin{array}{c|c} M_{11} & M_{12} \\ \hline M_{21} & M_{22} \end{array}\right)\left(\begin{array}{c} \iota_3 \\ \iota_2 \end{array}\right) = \left(\begin{array}{c} M_{11}\iota_3 + M_{12}\iota_2 \\ \hline M_{21}\iota_3 + M_{22}\iota_2 \end{array}\right) = \left(\begin{array}{c} \iota_3 \\ \hline \iota_2 \end{array}\right) \equiv \iota_5, \tag{49}$$

where ι_n is the all ones column vector of length n (see (8)).

Let $\bar{M}$ be the Markov matrix for the states A and B:

$$\bar{M} = \left(\begin{array}{cc} r_{11} & r_{12} \\ r_{21} & r_{22} \end{array}\right), \tag{50}$$

where the entries r_{ij} can be derived from (47) using the following system of linear equations:

$$\begin{aligned} r_{11} &= \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{3} p_{ij}, \\ r_{12} &= \frac{1}{3}\sum_{i=1}^{3}\sum_{j=4}^{5} p_{ij}, \\ r_{21} &= \frac{1}{2}\sum_{i=4}^{5}\sum_{j=1}^{3} p_{ij}, \\ r_{22} &= \frac{1}{2}\sum_{i=4}^{5}\sum_{j=4}^{5} p_{ij}. \end{aligned} \tag{51}$$

Rewriting (51) in matrix form yields

$$\bar{M} = DK'MK, \tag{52}$$

where

$$D = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}^{-1} \tag{53}$$

and

$$K = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \tag{54}$$

So the linear transformation $\mathcal{T}$ defined by

$$\mathcal{T}(M) = DK'MK = \bar{M}, \tag{55}$$

describes the aggregation of a 5×5 Markov matrix M into a 2×2 Markov matrix $\bar{M}$. □

Now we turn to the case important for the present paper: transition from cell densities to municipal densities. We start with a Markov matrix M_C^h derived by the telecom provider from the cell phone data, at the cell level. We group the cells on the basis of the municipalities in which they are located in (each cell is located in exactly one municipality). As in the case of the small example above we compute M_μ^h, the aggregated Markov matrix at the municipal level. Below we give the details by involving the various densities and transformations. If we use the same notation as in the example above to denote this transformation, i.e. $\mathcal{T}$, we have

$$\mathcal{T}(M_C^h) = DK'M_C^hK = M_\mu^h, \tag{56}$$

where K is similar to (54) and is a matrix that defines which cells are located in which municipalities. D is a matrix similar to (53) and which is the inverse of a matrix with the number of cells per municipality on the main diagonal.

It should be stressed that $\mathcal{T}$ is independent of h, provided the cell locations as well as the definitions of the municipalities do not change during the observation period, which we had assumed from the outset.

Another point to note is that $\mathcal{T}$ is a transformation of transformations, namely the M_C^h. That distinguishes it from the other density transformations considered in the present paper.

14 Transformations in context

The aim of the present section is to bring together all the densities considered in this paper so they can be studied in context. Table 14.1 presents an overview of all the linear transformations introduced in the present paper, with defining equations and references to the formulas in the text where they have been defined.

Transformations	Equations	Formulas
$\mathcal{G}$	$f_{GC}^h = \mathcal{G} f_C^h$	(14)
$\mathcal{V}$	$f_V^h = \mathcal{V} f_C^h$	(20)
$\mathcal{M}$	$f_\mu^h = \mathcal{M} f_C^h$	(33)
$\mathcal{S}$	$f_{SV}^h = \mathcal{S} f_C^h$	(36)
M_C^h	$f_C^{h+1} = \left(M_C^h\right)' f_C^h$	(37)
$\mathcal{M}_{GC}^h, \mathcal{G}$	$f_{GC}^{h+1} = \mathcal{M}_{GC}^h f_C^h = \mathcal{G} f_C^{h+1}$	(40), (14)
$\mathcal{W}_V^h, \mathcal{V}$	$f_V^{h+1} = \mathcal{W}_V^h f_C^h = \mathcal{V} f_C^{h+1}$	(45), (20)
$M_\mu^h, \mathcal{M}$	$f_\mu^{h+1} = \left(M_\mu^h\right)' f_\mu^h = \mathcal{M} f_C^{h+1}$	(46), (33)
$\mathcal{T}$	$M_\mu^h = \mathcal{T}(M_C^h) = DK' M_C^h K$	(56)
$\mathcal{R}$	$\tilde{f}_C^h = \mathcal{R}' f_C^h$	(B.7)

Table 14.1 The linear transformations defined in this paper.

It is clear from Table 14.1 that the cell densities f_C^h play a key role in the other densities that have been introduced in the present paper. It is a nice feature of the approach that geometry/geography and statistics can be cleanly separated. For some transformations in Table 14.1 (to wit $\mathcal{M}_{GC}^h$, $\mathcal{W}_V^h$ and M_μ^h) this is not the case because these operators combine both parts, to stress that the transformations from f_C^h are linear. But they can be written as products of linear transformations as was shown in previous sections.

Also, if the location of cells does not change, these transformations do not change either. We assume this to be the case in the present paper. If the geometry/geography changes at each level[35] then the linear transformation would change accordingly. But this process is unrelated to the process that drives the changes of the cell densities.

In Figure 14.1 a digraph is shown how the linear transformations $\mathcal{G}$, $\mathcal{V}$, $\mathcal{M}$ and $\mathcal{S}$ are interrelated. They are about changes in the geometry/geography and do not involve any statistics, which is entirely represented by f_C^h. It should be remarked that f_{SV}^h suggests one particular smoothed version of a Voronoi density, but in fact several choices are possible. The densities in Figure 14.1 apply to a single hour block.

Multiplying the transformations in Figure 14.1 by f_C^h yields a digraph with the dependencies among the densities f_{GC}^h, f_V^h, f_μ^h and f_{SV}^h, which is depicted in Figure 14.2. The densities f_μ^h and f_{SV}^h, are the important ones. The other two, f_{GC}^h and f_V^h are intermediate ones, auxiliary to producing the important ones.

As time development of all the derived densities concerns, they all are derived from that of the cell density:

[35] Starting with the location of the cells, which in turn has an effect on the Voronoi partioning, which in turn affects the way they intersect with the municipal partitioning and also the smoothing of the Voronoi densities.

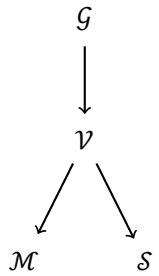

Figure 14.1 Interrelationship of transformations of densities for the same hourblock.

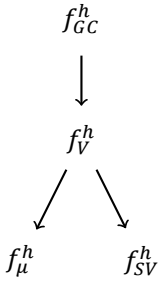

Figure 14.2 Interrelationship of densities for the same hourblock h. They are obtained from Figure 14.1 by multiplying each transformation with f_C^h.

$$f_C^{h+1} = \left(M^h\right)' f_C^h \tag{57}$$

Now (57) holds for $h = 1, \ldots, 23$ in our case. In Figure 14.3 the dependencies of the cell densities is depicted, including the transformations like (57) linking them. These conditions are ideal, as if no measurement errors are involved. In practice they will hold only approximately. These constraints then act as real constraints for the data which may have to be modified to satisfy these constraints. In Appendix E is discussed how this could be done.

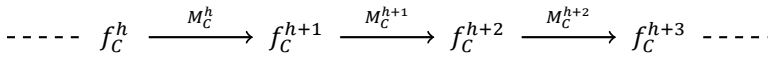

Figure 14.3 A sequence of cell densities and the Markov matrices linking them.

Once the f_C^h and M_C^h, for $h = 1, \ldots, 23$, are made consistent, so will be all the derived densities: f_{GC}^h, f_V^h and f_μ^h.

The linear transformation $\mathcal{R}$ is set apart from the other ones in the present paper, as it only uses geometric information, i.c the Voronoi polygons with the cell locations as generators, as auxiliary information to find cells closest to a given cell, forming a neighbourhood. The cell densities in each neighbourhood are averaged in some way (how, is not that important for the present, general discussion) and are assigned to the center cell of the neighbouring cells.

15 Animated density flow

In Section 7.3 we considered the graphical presentation of smoothed Voronoi density for an hour block h as a heatmap. This is a static situation. If one is interested in the change of the population density over time, one should show a sequence of such densities, in quick succession, like a film. This 'film' shows directly, and immediately understandably, how population densities change over time. The presentation is aimed at visualizing change of density and not about static density. It should be borne in mind that the density estimates are based on active cell phones. So during the night there will be little use of cell phones. This would (incorrectly) suggest low densities. But looking at the dynamics, it would (correctly) suggest small changes in density.

The smoothed Voronoi densities are used only to give a graphical idea of the dynamic population densities, by rendering them on a map, using different colours and shades of colours to show differences in densities locally. The contours of the municipalities are plotted in such maps as well only to provide some visual support for the user. The municipalities thus play a somewhat different, more passive, role than in the numerical presentation.

In principle the original Voronoi densities could be used as well, saving the effort to compute the smoothed Voronoi densities. In case they prove to be aesthetically less pleasant, one can still decide to compute the smoothed densities.

16 Helmholtz flow decomposition

The main goal of this paper is to derive population densities per hour block, starting with the cell level and proceeding to the municipality level. But this only yields separate population densities, without a clue about how they evolve from each other. So the next step is to study the development of these densities. In previous sections we considered the population flow between cells, due to moving cell phones and hence users. We could leave it at this. Or we go one step further and analyze this density flow. This can, for instance, be done by computing the Helmholtz decomposition of the flow, except the loops. This actually means leaving out the main diagonal entries in Markov matrices and concentrating on the off-diagonal entries, as will be explained.

In the networks considered in [14] no node has any capacity to hold 'mass'. 'Mass' that flows in a node also flows out at the next occasion. This is in contrast with the application studied in the present paper: the cells have loops (and hence also the municipalities have loops). This indicates that cell phones active in a particular cell c in hour block h may also be active in cell c in the next hour block $h + 1$.

In fact, loops indicate local 'traffic', indeed very local traffic. We can separate those loops immediately from the flow and concentrate on its remainder. For this remainder Kirchhoff's law holds. And to this remainder we can apply Helmholtz flow decomposition as decribed in [14]. We do not wish to discuss this subject here, because it can be investigated best when flow data are available.

17 Discussion

In the present paper an approach is sketched to estimate population densities for hour blocks, that is time intervals of one hour. Also consideration is given to the change of these densities from one hour block to the next. The data used for this model is cell phone data, from anonymous cell phones. Any cell phone that is within the reach of the network of cells in The Netherlands counts if it is active and provided that it has a subscription with the provider. This includes users from outside The Netherlands when they happen to be in The Netherlands. And it excludes Dutch cell phone users when they stay abroad.

With such data it is only possible to estimate the presence of cell phones in each hour block. It is not possible to track any cell phone over a longer time. If the goal was only to estimate hourly densities this would be sufficient. But we also want to say something about how densities change. To do this, extra information is required, which provides information about transitions made from one hour block to the next. This is comparable to a road network, where information is given about how many cars go left, straight-on, or turn right, at certain time intervals. This is traffic information which is aggregate information. It does not contain information about individual cars. So realistically, it is impossible to follow any car in the network.

In our case, cell phone location is in terms of cells. Cells are the objects through which cell phones communicate with each other, with landline phones, with webpages on the internet, etc. For our model it is only important when cell phones are active, and with which cells they communicate (and for how long at which cells) when they are active. For simplicity we can assume that an active cell phone communicates through the nearest cell in the network. But the model can still be used if there is a switch of cells during a session of a cell phone in active use, provided the presence at each of the cells involved can be computed.

In contrast with cars that physically do not appear or disappear suddenly[36] active cell phones do not have this persistence: the users of these cell phones can switch them on or off if and when they want to. Also cell phones can enter the network, when they enter the country, or they can leave it, when their users decide to.[37] So the collection of cell phones is an open population.

In terms of flow of ‘mass’ through the cell network, a drawback of this openness is that Kirchhoff’s law concerning mass preservation does not hold. This makes it difficult to describe the (anonymous) flow of ‘mass’ through the network. For that reason we do not consider the number of (anonymous) units flowing through the network. Instead, we describe the change of population densities.

Due to the possibility cell switching (irrespective of a cell phone moving or not) implies that the presence of a cell phone is not necessarily linked to a single cell, but may be associated with several cells during an hour block. This in fact means that cell switching tends to diffuse the location of cell phones. And in the next hour block the same may happen. So one ‘distributed presence’ at some hour block may change into another one at the next hour block.

[36] Which would be different if they had a device (a tracker) which sends its position at regular time intervals, which would be possible to switch on or off.

[37] This property they have in common with cars.

So at the cell level we have hourly cell densities as well as Markov matrices providing information about how these densities change, that is, their dynamics. If we associate with each cell location the density of an hour block, we have made the first step towards density visualization on a map. We call this the geometrical cell density.

The next step is to produce from this geometrical cell density a Voronoi density, that is a density based on the Voronoi partition generated by the cell locations.[38)] The underlying idea is that the density value at each cell location (for an hour block) is evenly distributed over the corresponding Voronoi polygon. This gives useful but somewhat crude density information. We can then proceed in two ways. In the first one we translate the density information for a Voronoi partitioning generated by the cell locations into another partition, one which is of relevance in official statistics, such as one based on municipalities. Each Voronoi polygon that intersects a municipality donates a part of its density mass to that of this municipality. Municipalities form an example of a geometric partitioning of the country that makes sense geographically and statistically, in contrast to the Voronoi partition based on the cell locations. In the second way to proceed with a Voronoi density is to smooth it and then colour code the resulting density, thus obtaining a smoothed Voronoi density. By playing these visualized densities like a film one can visualize their change.

As to the dynamics of the cell densities we have Markov matrices for each hour block pair $(h, h+1)$ in the observation period. We can use these to compute the Markov matrices to describe the transitions between municipalities, because it is known which cells are located in which municipalities. So the Markov matrices for municipalities can be obtained by aggregating Markov matrices at the cell level, by grouping the cells located in the same municipality.

It turns out that all the transformations used in the present paper, from one density to another or from one Markov chain to another, are all linear, which is pleasing and interesting. It indicates an inherent simplicity of the approach.

This is in a nutshell how the method is supposed to work. But there are a few problems that have to be looked at carefully. And this should be done, ideally, with the use of real cell phone data. The present paper is purely theoretical, and whose aim was to sketch an approach. The next step should be to implement the method proposed and see how well it works in practice. There are several issues that have to be addressed. They will be discussed next.

The first issue we consider concerns the basic idea of using cell phone data for estimating dynamic cell densities per hour block. How suitable are such data to yield good esimates of the real population densities per hour block? Is the group of persons using cell phones representative of the entire population? For sure it is not representative of very young and very old people, as they tend to not use cell phones at all, or in a relatively small percentage of their respective age groups. Elderly people tend to be less mobile than other age groups. Very young children tend to be near to (at least one of) their parents.

The next issue that we want to bring forward is the fact that the method is based on active cell phones. This creates a problem when many people are not actively using their cell phones, for

[38)] This is in a basic model where we are not supposed to have more technical information about the cells in the network, in particular about their directionality or about their sensitivity due to objects in their vicinity. If this information is actually available more realistic models can be developed and applied, as will be indicated below.

instance during the night, when they are asleep. During the day inactivity of cell phones is more likely to occcur at any moment, and for any user, but not as massively as during the night. Therefore this is likely to be less of an issue then.

Perhaps a slightly different approach from the one taken in this paper could be of help, namely one that starts with a population density at some point in time. And observations of active cell phones are used to estimate density changes as time progresses.

In fact, using models to estimate the location of previously active cell phones that have been switched off in some hour block, on the basis of observed mobility, is another interesting and worthwhile problem to study. The same is true for cell phones that have been switched on or have entered the country.

In the approach presented only billable information was used, in particular about the time used for various services. This information was used to compute the presence (i.e. q-presence) at cells. The users who answer a call are not taken into account.[39] The question is if another, simpler approach is feasable and attractive, namely one in which all connections made by cell phones with cells in the network are used, irrespective of whether they are at the sending or receiving end of calls, that is, whether they concern making calls or being called. Of importance is only when a link between a cell phone and a cell is made.

In such an approach it may also be interesting to investigate the possibility to ignore the presence[40] of cell phones and only use that a cell phone id in hour block h was connected to cells c_1, c_2, …, c_{k_h}. Each cell can appear at most once in hour block h. The number $n_{c,h}$ of different cell phones connected to cell c and for hour block h from the basis of the cell frequency of hour block h in this approach. The cell density flow could be handled in a similar way as in the present paper.

When several location methods of cell phones have been implemented it is possible to compare the resulting estimates of dynamic population densities, investigate possible differences, and identify the best method among the competing ones.

In the present paper we assume that there was only one telecom provider supplying all cell phone data. In practice it is probably preferable to use data from several telecom providers. These data may then be pooled (after some preprocessing) thus increasing the volume of data to be used. And the results for the various providers may be compared. And it will become clear whether different telecom providers have different types of customers with different use of their smartphones, different mobility patterns and different types of places that they tend to spend their time.

Another issue concerns the Voronoi polygons. Above it was suggested that always entire polygons are used in the computations. This, however, is only the case for interior cells. For boundary cells the situation is somewhat different. And it even matters whether these boundary cells are terrestrial cells or near the coast. And it may be necessary to use boundary cells from neighbouring countries (Belgium and Germany, in the case of The Netherlands, which was taken as the example country in the present paper). Boundary cells near the coast are treated

[39] It was argued in Section 3.1 that in case of asynchronous communication (e.g. text messages) this is not a problem.
[40] As defined in Section 2.4.

differently, namely by truncating them and consider only the parts that cover land. The idea is that the parts that cover water (the sea, actually) contain very few, if any, (active) cell phones.

Yet another subject for future research concerns the length of the time blocks. We have chosen hour blocks in the present paper, following [2]. They seem to be rather long. It is of interest to investigate whether shorther time blocks are feasible, particular in view of possible compromise of privacy. Of course, it is also the question how much detail is needed for statistical purposes. There is no sense in using time windows that are too short. So the question arises which size of the time window is optimal.

It should be stressed that a key characteristic of the model used is that actual routes followed by cell phones (at the cell level) cannot be used, due to the anonimity of the cell phones. There is a shift from individual cell phones to a collective of anonymous cell phones. The flow of this 'mass' is described by an inhomogeneous Markov chain, with a Markov matrix for each pair of consecutive hour blocks. This process is comparable to generalized diffusion process, where transition probabilities may change over time. As hour blocks are used, each moving and active cell phone is represented by a density of cell locations. The total presence per cell is the sum of the presences of all the cell phones active in that cell at a particular hour block. Evidently it is impossible to identify any cell phone (and hence its user).

In the approach taken in the present paper it was assumed that no specifics about the cells are known concerning their sensitivity. This was done for simplicity, to create a base model that can be applied straightforwardly. However, if information about the sensitivity of cells is available, in particular about the sensitivity areas of each cell, the model could be adapted to this situation. How this can be done is illustrated in Appendix F.

Another line of research would be to investigate whether data on cell phones that are switched on but not necessarily active could be used to produce estimates of population densities. Whether such data will be made available by telecom companies remains to be seen. These data are also not perfect. The risk using them is that a cell phone and its user happen to be at different locations. Think of a person who, in the evening, quickly visits a supermarket in his neighbourhood, leaving his cell phone at home. The question is, if it happens a lot that a cell phone and its user are separated, and also how much they are separated. But when asleep it is likely that the separation between the two is small. So for the evening such data would probably be great in estimating the locations of smart phones. They are also very useful to check if these data, as well as the data derived from active cell phones, give similar density estimates. They have the great advantage that they also provide location information for cell phones that were not active, but only switched on. And they can be used to distinguish for border cells, whether a cell phone has been switched off or on, or whether the cell phone left or entered the country. It is not necessary to use these data in the original frequency that they were generated. The frequency could be a lot lower and then would still produce useful information.

References

[1] V. Belikov, V. Ivanov, V. Kontorovich, S. Korytnik & A. Semenov (1997). The non-Sibsonian interpolation: A new method of interpolation of the values of a function on an arbitrary set of points. *Computational Mathematics and Mathematical Physics*, 37 (1), 9–15.

[2] CBS (2020). Estimating hourly population flows in the Netherlands, Discussion paper, Statistics Netherlands, Heerlen.

[3] N. Christ, R. Friedberg, T. Lee (1982). Weights of links and plaquettes in a random lattice. *Nuclear Physics B.*, 210 (3), 337–346.

[4] N. Cressie (1993) *Statistics for Spatial Data*. Wiley.

[5] E. de Jonge, M. van Pelt & M. Roos (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data, Discussion paper, Statistics Netherlands, The Hague & Heerlen.

[6] R. Little & D. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

[7] N. Mushkudiani & J. Pannekoek (2019). Estimating a time series of temporary employment using a combination of survey and register data, Discussion paper, Statistics Netherlands, The Hague.

[8] A. Okabe, B. Boots & K. Sugihara (1992). *Spatial Tessellations – Concepts and Applications of Voronoi Diagrams*. Wiley.

[9] F. Preparata & M. Shamos (1985). *Computational Geometry – An Introduction*. Springer.

[10] R. Sibson (1981). A brief description of natural neighbour interpolation. Chapter 2 in V. Barnett (ed.). *Interpreting Multivariate Data*. Wiley, 21–36.

[11] A. Strang (2020). Applications of the Helmholtz-Hodge Decomposition to Networks and Random Processes, PhD dissertation, Division of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University.

[12] D. Widder (1975). *The Heat Equation*. Academic Press.

[13] L. Willenborg (2017). From GEKS to cycle method, CBS, The Hague.

[14] L. Willenborg (2023). Helmholtz decomposition for digraphs, Discussion paper, CBS, The Hague.

Appendix
A Fictitious example data

In this appendix we present (entirely fictitious) example data to illustrate some points in the text. The data are presented in Table A.1. It is a table where the rows consist of records with information about clients/users (column id_i) that have been active on their phone in hour block h or $h+1$ when their signals were picked up by cells (c_j) and spending a certain amount of time ($\mathrm{p}_{i,j}$, which is a fraction of an hour). These data are supposed to be produced from even more basic data, source data generated from elementary events on the mobile telephone network: when an activity started, when it finished, which cells were involved. The data in Table A.1 are supposed to be derived from the source data by the telecom provider. They are not supposed to be shared with anybody else, in particular the statistical office. These data are sensitive, since they potentially could be linked to individuals. They serve in turn as a source of data for the statistical office: about densities per hour block, about linking of cells, and about the flow of 'population mass' through the cells network. In Sections 4, 8 and 9 it is shown how these data are used for various computations concerning dynamic population densities.

The variables 'cell' and 'id' can be viewed as secondary keys (in the language of relational database theory). This means that there are separate tables with details about each cell (like location, and technical information) and about each client (like name, address, bank account number, etc). But this information is of no use here. In fact the id's used are only needed to combine information from the same users. The id's should distinguish them. Therefore they need not be the original keys but they can be surrogate keys.[41)] In fact, the keys 'id' need not be the same throughout but they can depend on the pair h and $h+1$ of hour blocks. To stress this we could write $\mathrm{id}_i^{h,h+1}$ to indicate the dependence of the keys (identities) on the pair of consecutive hour blocks h and $h+1$. But as this involves purely the internal working of the telecom provider, we abstain from providing such details. They are not necessary to produce the data we are interested in. They are only an extra precaution for the safe internal handling of these data by the telecom provider. Note that with such (surrogate) keys it is impossible to track users for periods longer than two hours. If the telecom provider would take the trouble of using such dedicated surrogate keys, they would only increase the internal safety of the internal handling of sensitive data. It would be good general policy of the telecom provider if access to client data is only granted to employees who need them for their work. Data that employees do not need for this purpose should not be accessible to them.

Table A.1 is complete as to the activities on the cell phones in hour blocks h and $h+1$. However, not all cell phones were used in these hour blocks. In hour block h cell phones with the identities id_2, id_4, id_7, id_9 and id_{10} were not active. In hour block $h+1$ the same is true for the cell phones with the identities id_1, id_4, id_7, id_9 and id_{10}. We see different patterns:

- id_1 is active in hour block h but not in hour block $h+1$.
- id_2 is active in hour block $h+1$ but not in hour block h.
- id_3 is active in both hour blocks h and $h+1$.

41) If Ω is the set of original keys and Σ is a set of surrogate keys then $|\Omega| = |\Sigma|$, so they are equal in size, and there is a bijection $\kappa : \Omega \to \Sigma$.

– id_4 is not active in both hour blocks h and $h+1$.

rec	h bl	cell	id	q-pr	rec	h bl	cell	id	q-pr
1	h	c_1	id_1	0.67	21	h+1	c_6	id_{11}	0.78
2	h	c_2	id_3	0.19	22	h+1	c_6	id_{12}	0.15
3	h	c_2	id_5	0.08	23	h+1	c_7	id_{12}	0.48
4	h	c_3	id_5	0.25	24	h+2	c_2	id_1	0.38
5	h	c_3	id_6	0.33	25	h+2	c_2	id_2	0.21
6	h	c_4	id_6	0.21	26	h+2	c_3	id_4	0.19
7	h	c_4	id_8	0.14	27	h+2	c_4	id_5	0.41
8	h	c_5	id_8	0.13	28	h+2	c_5	id_6	0.14
9	h	c_6	id_8	0.21	29	h+2	c_5	id_7	0.12
10	h	c_6	id_{11}	0.49	30	h+2	c_7	id_8	0.13
11	h	c_6	id_{12}	0.16	31	h+2	c_7	id_9	0.63
12	h	c_7	id_{12}	0.71	32	h+2	c_1	id_9	0.26
13	h+1	c_1	id_2	0.18	33	h+2	c_1	id_{10}	0.47
14	h+1	c_2	id_2	0.63	34	h+2	c_2	id_{11}	0.19
15	h+1	c_3	id_3	0.72	35	h+2	c_3	id_{11}	0.22
16	h+1	c_3	id_5	0.23	36	h+2	c_4	id_{11}	0.23
17	h+1	c_4	id_5	0.29	37	h+2	c_5	id_{12}	0.11
18	h+1	c_4	id_6	0.13	38	h+2	c_6	id_{12}	0.16
19	h+1	c_4	id_8	0.24	39	h+2	c_7	id_{12}	0.15
20	h+1	c_6	id_8	0.11					

Table A.1 Records (rec) of active cell phones (id) at cells (cell) in three consecutive hour blocks (h bl) and their q-presence (q-pr).

From these data the telecom provider can compute, for each client, the total amount of time they spent using the telecom facilities per hour block. This is important billing information for the telecom provider, but it is of no importance for the application that we are interrested in in the present paper.

Instead for our application the q-presence (q-pr) is important. In each record in Table A.1 a q-pr is linked to a cell in an hour block (h or $h+1$). It is possible to compute a few quantities that are of interest to our approach:

1. total q-presence per cell.
2. cell link data.
3. cell flow information.

We now define each of these variables. Let the q-presence of (q-pr) cell phone id_j at cell c_i in hour block h be denoted by $\text{q-pr}(c_i, h, \text{id}_j)$ and the total q-presence in hour block h at cell c_i by $\tau(c_i, h)$. Then

$$\tau(c_i, h) = \sum_{\text{id}_j} \text{q-pr}(c_i, h, \text{id}_j). \tag{A.1}$$

So for instance, if we consider cell c_2 and hour block $h+2$, then we find, using Table A.1, that cell phones id_1, id_2 and id_{11} were active, so that the total presence equals $\tau(c_2, h+2) = 0.38 + 0.21 + 0.19 = 0.78$.

The cell link data are about the interconnection of cells as a result of moving cell phones that are active in consecutive hour blocks h and $h+1$. We again derive this information from Table A.1. In fact, this information leads to a digraph, the cell link digraph, in which the nodes are the cells and links denote ordered pairs of cells (c,d) such that there was at least one cell phone id_k and hour block h such that id_k was active at cell c in hour block k and in cell d in hour block h or $h+1$. What matters is the fact that d can be reached from c in the same hour block or the one immediately following. Unimportant here is how long id_k was active in cell c or d, as long as this was time > 0 (or, alternatively, q-pr > 0) for both cells. For instance cell phone id_5 in Table A.1 is active at cells c_2 and c_3 in hour block h and in cell c_3 and c_4 in hour block $h+1$. Therefore we have the arcs $\{c_2,c_3\}$, (c_2,c_4), (c_3,c_3), $\{c_3,c_4\}$ for the link digraph in our example.[42)]

Note that we have assumed that cell c_3 can be reached from c_2, and vice versa, in one hour block. The same is true for cell c_3 and c_4. Strictly speaking we cannot deduce this from the data as we have no information about the order in which cell phone id_5 was active in the cells. It is possible that c_2 can be reached from c_3 within one hour block (or also including the next one) but not the other way round. But as we have no information about this, we assume that the most optimistic scenario about the reachability of cells is actually the case.

Note also that we have generated the loop (c_3,c_3). This is only because the cell c_3 was active in different hour blocks, i.c. h and $h+1$.

Finally we consider the cell flow information. In this case the presence of cell phones in cells is also of no importance, except that both should be nonnegative. We take an arc (c,d) of the cell link digraph that we have just considered. We now look for all cell phones id_k in Table A.1 that are active at cell c in hour block h and also active at cell d in hour block h or hour block $h+1$. Then a measure of the size of the flow from c to d is the number of such cell phones, making the move between hour block h and hour block $h+1$. We can represent this flow information by putting these flow numbers $\nu_{cd}^h \in \mathbb{N}^0$ as tags to the corresponding arcs (c,d). They can be stored collectively in a flow matrix

$$\Phi_C^h = \begin{pmatrix} \nu_{11}^h & \cdots & \nu_{1n}^h \\ \vdots & \ddots & \vdots \\ \nu_{n1}^h & \cdots & \nu_{nn}^h \end{pmatrix}. \tag{A.2}$$

From matrix (A.2) we can produce the Markov matrix

$$M_C^h = \begin{pmatrix} \nu_{11}^h/\nu_{1\cdot}^h & \cdots & \nu_{1n}^h/\nu_{1\cdot}^h \\ \vdots & \ddots & \vdots \\ \nu_{n1}^h/\nu_{n\cdot}^h & \cdots & \nu_{nn}^h/\nu_{n\cdot}^h \end{pmatrix} = \begin{pmatrix} f_{11}^h & \cdots & f_{1n}^h \\ \vdots & \ddots & \vdots \\ f_{n1}^h & \cdots & f_{nn}^h \end{pmatrix}, \tag{A.3}$$

where $\nu_{i\cdot}^h = \sum_{j=1}^n \nu_{ij}$ and $f_{i\cdot}^h = 1$, for $i = 1,\ldots,n$. See also Section 9.1.

[42)] It should be stressed that $\{a,b\} = \{(a,b),(b,a)\}$.

B Geometric smoothing

In this appendix we discuss some methods that can be used to produce Voronoi densities from cell densities, where cells are represented as dots in a map. Or they can be used to smooth cell densities. Or to smooth Voronoi densities, in case they are used to produce pictures of densities that hide the details of Voronoi partitions and thus provide a more gently, flowing picture. We also remark that this smoothing is a linear transformation when applied to densities on maps, in which we are particularly interested.

B.1 Nearest-neighbour interpolation

Suppose that n points $p_1, \dots, p_n$ in the plane are given, and values $y_1, \dots, y_n$ in $p_1, \dots, p_n$, respectively. A simple interpolation method is based on the Voronoi partition induced by the p_i. For Voronoi polygon V_{p_i} associated with p_i we assume that the value of the interpolant is constant and equal to y_i. So clearly the interpolating function is locally constant (on each Voronoi polygon) and is likely to be discontinuous if neighbouring Voronoi polygons have different associated y-values. So in general we may expect an interpolant that is not continuous.

A variant of nearest-neighbour interpolation that we use in the present paper to obtain Voronoi densities from cell densities is as follows. The value for Voronoi polygon $V(p_i)$ associated with p_i, the location of cell c_i, is $y_i/|V_{p_i}|$ (instead of y_i), where $|V_{p_i}|$ denotes the area of V_{p_i}. Thus the 'density mass' y_i is preserved, being evenly distributed evenly over V_{p_i}. See Section 6.2. Of course, one can view this operation as a combination of two operations: nearest neighbour interpolation followed by an adjustment for each value associated with a Voronoi polygon.

B.2 Natural neighbour interpolation after Sibson and after Laplace

The form of the interpolating function $G : \mathbb{R}^2 \to \mathbb{R}\backslash\mathbb{R}^-$ that we consider here is as follows:

$$G(x) = \sum_{i=1}^{n} w_i(x) f(p_i), \tag{B.1}$$

where $p_i \in \mathbb{R}^2$ for $i = 1, \dots, n$.

Natural neighbour interpolation (NNI)[43] is a method to create smoother functions than nearest neighbour interpolation produces. NNI in turn is a natural extension of this interpolation method. For each point x for which one wants to compute the value of the interpolated function one adds x to the list of generators $\ell(p)_i$, for $i = 1, \dots, n$, where ℓ is as defined in (10). This creates a Voronoi polygon V_x associated with x, which intersects with some of the Voronoi polygons of the original Voronoi partition.

[43] The discussion presented here is based on `https://en.wikipedia.org/wiki/Natural_neighbor_interpolation` and the pictures shown are also taken from this site.

In Figure B.1 the green coloured circular areas represent the interpolation weights. The purple-shaded region is the new Voronoi polygon, after inserting the point to be interpolated (the black dot). The weights represent the intersection areas (relative to the area of V_x) of the purple-polygon with each of the seven surrounding cells. The value of the interpolating function at x is the weighted sum of the values of the original function at the original generators of the Voronoi partition. This interpolation method is called 'Sibson interpolation', after the author of [10], where this method is proposed.

Symbolically a Sibson weight can be expressed as

$$w_i^S = \frac{|V_{p_i} \cap V_x|}{|V_x|}, \tag{B.2}$$

where $|\cdot|$ denotes the area function.

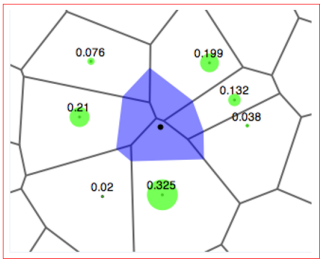

Figure B.1 Natural neighbour interpolation with Sibson weights.

The weights used for interpolating can also be obtained in a different way, which is possibly computationally more efficient, namely by using the distances $d(x, p_i)$ of x to each of generators of the Voronoi polygons with a non-empty intersection with V_x. In Figure B.2 the interface between the polygons linked to x and p_i is in blue (and of the length $l(p_i)$), while the line segment connecting x and p_i is in red (and of length $d(x, p_i)$). $l(p_i)$ and $d(x, p_i)$ can be used to compute so-called Laplace weights (cf. [1] and [3]). Symbolically a Laplace weight can be expressed as

$$w_i^L = \frac{l(p_i)/d(x, p_i)}{\sum_{k=1}^{n} l(p_k)/d(x, p_k)}. \tag{B.3}$$

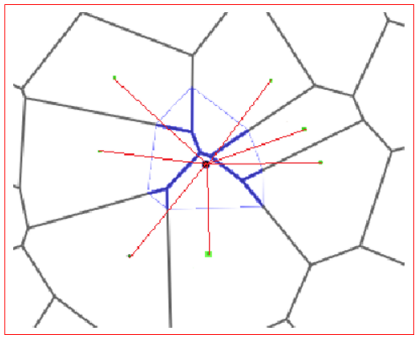

Figure B.2 Natural neighbour interpolation with Laplace weights.

B.3 Smoothing cell densities

In the present section we consider a different smoothing technique, not carried out on densities defined on geometrical structures like Voronoi polygons, but directly at the cell level. The methods developed in the present paper can then be used to represent these smoothed values. The Voronoi polygons generated by the cell locations are used to define neighboring cells. Let c be a cell and V_c its Voronoi polygon, then we define $\mathcal{N}_c^1$ as its (direct) neighborhood. It consists of Voronoi polygons V that border V_c, in the sense that V_c and V have a line segment[44] in common. This means that $V_c \cap V$ is a line segment. If the intersection consists of only a point then V_c and V are not bordering. The superscipt 1 applies to the fact that we are dealing with Voronoi polygons at distance 1, which means that they are direct neighbours. Using this notation we can define $\mathcal{N}_c^p$ with $p = 0, 1, 2, \ldots$. We define $\mathcal{N}_c^0$ to consist of V_C only. $\mathcal{N}_c^2$ consists of those Voronoi polygons that border the polygons in $\mathcal{N}_c^1$ insofar they are not in $\mathcal{N}_c^1$ or in $\mathcal{N}_c^0$. Likewise, $\mathcal{N}_c^3$ consists of Voronoi polygons bordering those in $\mathcal{N}_c^2$ insofar they are not in $\mathcal{N}_c^0$, $\mathcal{N}_c^1$ or $\mathcal{N}_c^2$. Likewise for larger values of p.

We can describe the neighbourhood structure on the set of Voronoi polygons by a $n \times n$ adjacency matrix $\mathfrak{A}_V = (\mathfrak{a}_{ij})$ or by an $n \times n$ incidence matrix $\mathfrak{I}_V = (\mathfrak{i}_{kl})$. The entries of $\mathfrak{A}_V$ are indexed by Voronoi polygons i and j. The entries of $\mathfrak{I}_V$ are indexed by neighbourhoods $\mathcal{N}_k$ and Voronoi polygons V_l. We have

$$\mathfrak{a}_{ij} = \begin{cases} 0 & : \text{ if } V_i \text{and } V_j \text{are not neighbours,} \\ 1 & : \text{ if } V_i \text{and } V_j \text{are neighbours.} \end{cases} \tag{B.4}$$

and

$$\mathfrak{i}_{kl} = \begin{cases} 0 & : \text{ if } V_l \notin \mathcal{N}_k, \\ 1 & : \text{ if } V_l \in \mathcal{N}_k. \end{cases} \tag{B.5}$$

It should be stressed that the adjacency matrix $\mathfrak{A}_V$ is different from that of the cell link digraph, as presented for a toy example in Table 8.5. The corresponding (di)graphs have the same set of vertices (namely the set of cells) but the set of arc/edges are defined differently.

Once the neighbourhoods for the Voronoi polygons have been identified, we consider the cells that are the generators of the Voronoi polygons in each neighbourhood. So if

$$\mathcal{N}_{V_c} = \{V_{c_{i_1}}, \ldots, V_{c_{i_t}}\} \tag{B.6}$$

is such a neighbourhood then we are interested in the set $\{c_{i_1}, \ldots, c_{i_t}\}$, the generators of the Voronoi polygons in $\mathcal{N}_{V_c}$. The idea is now to average the values of the vector of densities f_c^h for each of these neighbouring cells and assign the computed average to the c, the cell with neighbourhood $\mathcal{N}_{V_c}$. This averaging can be done in various ways. For instance one can use the

[44] Corresponding to an edge defining the boundary of Voronoi polygons.

idea underlying Kriging, a method used in geostatistics, where a convex combination of the values (in our case, of population densities for hour blocks) in a neighbourhood is taken, in such a way that its variance is minimized.[45] Or one could weigh each value by the reciprocal of its (estimated) variance. At any rate, this provides us with a Markov matrix $\mathcal{R}$ that describes the averaging per neighbourhood. A smoothed version of the density f_C^h for hour block h using this matrix is denoted by $\tilde{f}_C^h$, where

$$\tilde{f}_C^h = \mathcal{R}' f_C^h, \tag{B.7}$$

where ' denotes transposition, as elsewhere in this paper.

The methods described above can also be applied to $\tilde{f}_C^h$ instead of to f_C^h.[46] The smoothing using $\mathcal{R}$ can be repeated several, say k, times, if desired. We then obtain

$$\tilde{\tilde{f}}_C^{h,k} = (\mathcal{R}')^k f_C^h. \tag{B.8}$$

One would expect to choose small values of k, such as 2 or 3, to obtain double or triple smoothing. A priori there is no obvious argument why this multiple smoothing would be necessary. It is only mentioned to point out that it is possible.

B.4 Averaging over circular neighbourhoods of points

Apart from the interpolation method considered in Sections B.1 and B.2 we want to mention one here that is not particularly geared at Voronoi polygons but is of a more general nature. In a sense it is the embodiment of smoothing of a function f at a point x in its domain: replace the value $f(x)$ at x by the average of the values in a neighbourhood of x and divide by the size of this neighbourhood. As a neighbourhood we can take a disk $B_r(x)$ with radius $r > 0$ and center x, in case the domain of f is $\mathbb{R}^2$ or a subset L thereof. In order to avoid problem at points near the border of L, we consider $B_r(x) \cap L$. We then define a smoothed version of $f : L \to \mathbb{R}$ as

$$\bar{f}_r(x) = \frac{1}{|B_r(x) \cap L|} \int_{B_r(x)\cap L} f(y)\, d\mu(y), \tag{B.9}$$

where μ is a suitable measure on $L \subseteq \mathbb{R}^2$.[47] Of course, if x is sufficiently removed from the boundary of L we have $B_r(x) \cap L = B_r(x)$. If L is bounded and r is sufficiently big, namely such that $B_r(x) \supseteq L$, then $\bar{f}_r$ is a constant function.

It is clear that the effect of applying this smoothing is that it removes extremes in a function f. If applied repeatedly it will result in a constant function. If applied to a probability density it will

[45] The method is named after Danie Krige from South Africa. For a description of Kriging see e.g. [4].
[46] With the possible exception of the method for smoothing Voronoi densities, using the operator $\mathcal{S}$ (see (36)), if smoothing twice is not wanted.
[47] Such as a Riemann or a Lebesgue measure.

increase the entropy and produce a distribution $\bar{f}_r$ that is flatter and that tends to a uniform distribution (which exists, as L is bounded).[48)]

Clearly, (B.9) is a linear transformation. To stress that a linear transformation was used we write $\Xi_s f$ instead of $\bar{f}_r$.[49)] We have

$$\Xi_s(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 \Xi_s(f_1) + \lambda_2 \Xi_s(f_2), \tag{B.10}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$.

The interpolation methods presented in Sections B.1 and B.2 can be described as linear transformations of the input functions.

If we are dealing with densities it should be noted that they form a convex set and not a linear space: if f_1 and f_2 are densities and $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$ then $\lambda_1 f_1 + \lambda_2 f_2$ is also a density. (B.10) interpreted in this context shows that the smoothed version of a density that is a convex combination of two densities, is a convex combination of the smoothed versions of the component densities, with the same convex weights.[50)]

[48)] If the goal is to obtain a density from the smoothing of a density by applying (B.9) then a normalisation may be necessary, so that the integral (or sum) of the resulting function over its domain equals 1.

[49)] This notation discards the parameter r that has been used, not to mention L, the distance function used to measure the radius of the disk, the measure used for integration, that all play a role. In the present context these things are details.

[50)] If f is a density Ξf may not be, as all its value added up (or integrated over its demain) need not be unity. We assume that this total (or integral) is finite and be equal to$N(\Xi f) < \infty$. Then $\Xi f / N(\Xi f)$ is a density.

C Diffusion

Through the process of diffusion we can get more insight into the spread of the cell phones over the country during the observation period. We only observe cell phones changing location via the cells in which they are active. We consider diffusion processes for continuous and discrete space and continuous and discrete time (four combinations in total). In our case we are dealing with discrete time (hour blocks) and discrete space (defined by the cells). Via diffusion there is a link with heat in physics (i.c. thermodynamics).

The purpose of this appendix is to link the problem of cell phone spread with a comparable phenomenon studied in physics. Heat diffusion in a piece of matter and spread of cell phones over an area can be viewed as similar phenomena, at least superficially. Looking at both processes more closely, there are differences: the cell phone population is open, whereas heat is preserved in a closed system.[51)]

We first look at the traditional physical setting, which uses continuous time and continuous space, then at the setting with continuous time and discrete space, to arrive at the setting where both parameters are discrete. This later setting is suitable for the problem considered in the present paper.[52)]

The data provided by the telecom provider only contains data on nonidentifiable (anonymous) cell phones. However we suppose it does contain 'link data', that is, data that indicate how active cell phones in the vicinity of a cell, in a particular hour block, may spread to neighbouring cells in the next hour block. This yields transition probabilities between cells, entries in Markov matrices. As these transition probabilities are likely to vary over time, we are dealing with nonstationary Markov chains and hence Markov matrices that are time dependent, that is, they depend on hour block h in our case.

C.1 Continuous space and time

We start with the equation that describes diffusion in case time and space are both continuous:

$$\frac{\partial u}{\partial t} = \text{div}\, D\, \nabla u = \nabla \cdot D \nabla u, \tag{C.1}$$

where $u(t,x)$ is a smooth function in time t and space $x = (x_1, x_2)$. In (C.1) 'div' is the divergence, ∇ the gradient and $D(t,x)$ a function that regulates the speed of the diffusion at time t and at location $x = (x_1, x_2)$. If the medium in which the diffusion takes place is

51) That is, if we consider equilibrium thermodynamics. Comparison with nonequilibrium thermodynamics is more apt. But this is a more complicated subject, that we wish to avoid. Dealing with it would miss the point of the present appendix: to draw attention to a classical subject in physics to develop a feeling for a key problem in the present paper.

52) The case with discrete time and continuous space would also be of interest if we would study the development of smoothed Voronoi densities. However, we only suggest to use these densities visually, not numerically, in the present paper.

homogeneous, D is a constant function with value $D_0 > 0$. In this case (C.1) reduces to the so-called heat equation:

$$\frac{\partial u}{\partial t} = D_0 \nabla \cdot \nabla u = D_0 \Delta u, \tag{C.2}$$

where $\Delta = \nabla \cdot \nabla$ is the Laplace operator. If we write (C.2) in terms of cartesian coordinates, we obtain

$$\frac{\partial u}{\partial t} = D_0 \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \right). \tag{C.3}$$

Equations (C.2) or (C.3) are linear partial differential equations. This means that if we have two solutions u_1 and u_2 so is any linear combination $\xi_1 u_1 + \xi_2 u_2$, for $\xi_1, \xi_2 \in \mathbb{R}$. We can interpret equations (C.2) or (C.3) as describing the spread of a particle determined by the coefficient D_0.

For more on the heat equation see e.g. [12].

For the present paper it seems to be appropriate to assume that D is not a constant function, but one that depends on time (t) but not on space x. The reason is that we have observations for transition probabilities for each hour block to the next. Dependence of D on x is not readily available and probably fairly complicated to determine or estimate.

C.2 Discrete space and continuous time

We now switch to discrete analogs of concepts of diffusion and heat exchange in Euclidean space ($\mathbb{R} \times \mathbb{R}^2 \simeq \mathbb{R}^3$) to those concepts defined on (di)graphs, where the nodes can be viewed as representing cells. For the moment, we still assume the time parameter to be continuous (that is, with values in $\mathbb{R}$).

We start with an example. In Figure C.1 part of a graph $G = (V, E)$ is shown. More in particular the neighbourhood of a vertex i and $j_1, \dots, j_5 \in V$ adjacent to i, that is, such that $\{i, j_1\}, \dots, \{i, j_5\} \in E$. We write $i \sim j_k$, for $k = 1, \dots, 5$, to indicate that i and the j_k are adjacent vertices. In the context of the application in the present paper this means that we are looking at cells i, j_k active in hour blocks h and $h + 1$, as a result of a moving cell phone that is active at cell i in hour block h and at cell j_k in hour blocks h or $h + 1$.

Now we consider the change of the value u_i at the red vertex i in Figure C.1. The changes in this vertex over time can only result from inflow from or outflow to adjacent vertices. If we assume that for each edge $\{i, j_k\}$ there is a parameter D_{ij_k} regulating the flow, such that $D_{ij_k} = D_{j_k i}$, we obtain

$$\frac{du_i}{dt} = \sum_{k=1}^{5} D_{ij_k}(u_i - u_{j_k}) = \mathcal{D}_i\, u_i - \sum_{k=1}^{5} D_{ij_k} u_{j_k}, \tag{C.4}$$

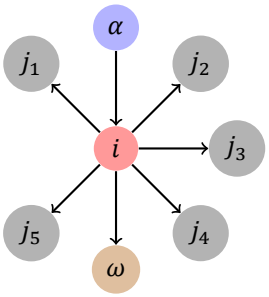

Figure C.1 A cell i (red) and the cells $j_1, \dots, j_5$ (black) that can be reached from it (leaving out a loop). The vertex α (blue) – the source – symbolizes the virtual location where cell phones come from when activated. The vertex ω (brown) – the sink – symbolizes the virtual location where cell phones move to when deactivated. In fact, α and ω could be identified.

where $\mathcal{D}_i = \sum_{k=1}^{5} D_{ij_k}$.[53)]

If we assume that $D_{i,j} = D_0$, a constant,[54)] for each edge $\{i, j\}$ then (C.4) simplifies to

$$\frac{du_i}{dt} = D_0 \sum_{k=1}^{5} (u_i - u_{j_k}) = D_0 \left(\deg(i)\, u_i - \sum_{k=1}^{5} u_{j_k} \right). \tag{C.5}$$

We may assume $D_0 = 1$ by suitably rescaling the time parameter, so that we finally arrive at

$$\frac{du_i}{dt} = \deg(i)\, u_i - \sum_{k=1}^{5} u_{j_k}. \tag{C.6}$$

We remark that an expression like (C.6) generalizes to

$$\frac{du_i}{dt} = \deg(i)\, u_i - \sum_{j \sim i} u_j, \tag{C.7}$$

which can be written more concisely as

$$\frac{du}{dt} = (\mathcal{D} - A)\, u = \mathcal{L}\, u, \tag{C.8}$$

where $u = (u_1, \dots, u_n)'$, $\mathcal{L}$ is the Laplace matrix,[55)] $\mathcal{D}$ the degree matrix, which is a diagonal matrix with the degrees of each of the nodes on the diagonal and the off-diagonal elements

53) We have tacitly assumed that each diffusion parameter D_{ij_k} is constant, to keep the derivation simple, for the purpose of illustrating the general approach, rather than a more accurate one.

54) This assumption is not necessarily realistic in view of our application, but rather it is convenient for our exposition on diffusion.

55) Or graph Laplacian

equal to 0, and A is the adjacency matrix of the graph. The Laplace matrix is symmetric, i.e. $\mathcal{L}' = \mathcal{L}$ and positive semi-definite $u'\,\mathcal{L}\,u \geq 0$ for all $u \in \mathbb{R}^n$.

The model based on (C.8) is used in [11], Chapter 6, equation (6.2). But, as already remarked, in our case this does not seem to apply. The diffusion function is not likely to be constant, but rather roughly cyclic, with a cycle of one day.

We can write the solution to (C.8) as

$$u(t) = \exp(\mathcal{L}t)\; u(0), \tag{C.9}$$

where

$$\exp(\mathcal{L}t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathcal{L}^k. \tag{C.10}$$

In an equilibrium situation[56] we have $\frac{du}{dt} = 0$, so that (C.7) implies that for each node i we have

$$u_i = \frac{1}{\deg(i)} \sum_{j \sim i} u_j. \tag{C.11}$$

Actually (C.11) expresses that the value of u at node i equals the average of the values of u in each node j adjacent to node i. Functions u with this property are called harmonic. In the continuous case harmonic functions u satisfy the classical Laplace equation

$$\Delta u = 0, \tag{C.12}$$

which explains the name of its counterpart in the discrete case.

A strict equilibrium state will never be reached (under normal circumstances). A state close to an equilibrium will be reached during the night when most people sleep and perhaps during the day when there are the most activities.[57]

[56] This may be a temporary equilibrium, or near-equilibrium.
[57] Such a state does not exist in case of a vibrant city, full of activity around the clock.

C.3 Discrete space and time

We now consider the final case, where both time and space is discrete. This is the case that is the most important one for the present paper. The space is is represented by the set of cells in the cell network. Time is discrete, or rather, discretized in hour blocks. As a cell phone may be on the move its signal, when active, may be picked up by several cells during different parts of an hour block. In that case the mass representing a cell phone during that time period[58)] is spread uniformly over these cells. This is assumed to be part of the telecom provider's intermediate data for the statistical office (see section 8).

The rest of the story is that of the dynamics of the cell densities and can be found in Section 9. From thereon several developments can be found in the main text.

58) The proportion of the time (on hour) it was active – its q-presence –, is distributed over the cells that picked up that cell phone's signal.

D Inactive cell phones and missing cell information

In the present appendix we want to consider the idea that inactive cell phones differ from active ones because their location (in the form of one or more cells) is missing. To remedy this the idea is to predict the missing locations by using a suitable model. These predictions can in turn be used to obtain new estimates for the cell densities f_C^h per hour block as well as new estimates for the Markov matrices M_C^h at the cell level.

We can illustrate our ideas with example data from Appendix A. From Table A.1 we derive Table D.1 with missing start or goal cells for hour blocks h and $h+1$. The statistical office only has the missing data patterns and their multiplicities, but is ignorant of the cell phones involved.

h	$h+1$	mult
c_1	-	1
-	c_1	1
-	c_2	1

Table D.1 Missing data patterns for hour blocks h and $h+1$ and their multiplicities (mult) and the cells (c_j) involved. It concerns cell phones active in exactly one of the hour blocks h or $h+1$.

$h+1$	$h+2$	mult
-	c_1	2
-	c_2	1
-	c_3	1
-	c_5	1
-	c_7	1
c_3	-	1

Table D.2 Missing data patterns for hour blocks $h+1$ and $h+2$ and the multiplicities (mult) and the cells (c_j) involved. It concerns cell phones active in exactly one of the hour blocks $h+1$ or $h+2$.

From Table A.1 we derive Table D.2 with missing start or goal cells for hour blocks $h+1$ and $h+2$. This table is similar to that in Table D.1. So these tables can be dealt with in a similar ways.

In case of Tables D.1 or D.2 it is possible to use the appropiate Markov matrices M_C^h, the Markov chain for the hour blocks h and $h+1$, to impute missing goal cells if the start cells are known. In case the goal cells are known we can use the reverse Markov matrices to impute the missing start cells. In both cases a missing at random assumption is used: the conditional distribution of cells in hour block $h+1$, given an observed cell in hour block h, is the same for cell phones that have been observed in both hour blocks as for cell phones that have been obseved only in hour block h. (And reversely for hour blocks h and $h+1$ interchanged.)

When we look at Table D.1 we see that we have to impute three records: For the first one we know that the start cell is is c_1. But we do not know the corresponding goal cell. To estimate this we can use the row of M_C^h that corresponds to this cell. We can then select a goal cell by drawing a cell using the probabilities in the row in M_C^h that corresponds to c_1. Likewise we can find the start cell that has cell c_2 as its goal. But then we first have to compute the reverse Markov matrix. Let M_C^h be the Markov matrix associated with the hour blocks h and $h+1$:

$$M_C^h = \begin{pmatrix} m_{11}^h & \cdots & m_{1n}^h \\ \vdots & \ddots & \vdots \\ m_{n1}^h & \cdots & m_{nn}^h \end{pmatrix}, \tag{D.1}$$

where the $m_{ij}^h \geq 0$ and the row sums are all equal to 1. The reverse Markov matrix $\tilde{M}_C^h$ is defined as follows

$$\tilde{M}_C^h = \begin{pmatrix} m_{11}^h/m_{\cdot 1}^h & \cdots & m_{n1}^h/m_{\cdot 1}^h \\ \vdots & \ddots & \vdots \\ m_{1n}^h/m_{\cdot n}^h & \cdots & m_{nn}^h/m_{\cdot n}^h \end{pmatrix}. \tag{D.2}$$

This matrix is the transpose of M_C^h, where also each entry in row i is divided by the column sum of column i of M_C^h. Note that the rows of $\tilde{M}_C^h$ all add to 1 and the entries are all nonnegative. Hence $\tilde{M}_C^h$ is a Markov matrix.

Referring to Table D.1 we can use $\tilde{M}_C^h$ to draw a cell that could have been the start cell for hour block h with c_2 as the goal for hour block $h+1$: use the row of this matrix as the distribution for the possible goal cells, and draw a cell. Impute this as the start cell with c_2 as the goal cell.

If we now look at Table D.2, we face similar problems as in case of Table D.1: five cells (c_1, c_2, c_3, c_5, c_7) for which start cells have to be generated and a single cell (c_3) for which a goal cell has to be drawn. This time we need to use M_C^{h+1} and $\tilde{M}_C^{h+1}$ instead of M_C^h and $\tilde{M}_C^h$, respectively. Note that the start cells also add to the cell density f^h. And, of course, these imputations imply a change of $\tilde{M}_C^{h+1}$ as well.

It should be stressed that the missing data problem that we consider here has to be dealt with for each consecutive pair of hour blocks, independently of each other. This may result in transitions that are incompatible in the sense that the goal cell of the first transition (concerning hour blocks h and $h+1$) is not the same as the start cell of the second transition (concerning hour blocks $h+1$ and $h+2$). We do not have the identities of the cell phones, so we may very likely create a situation that is different from the one originally observed: two transitions that cannot have been made by a single cell phone, as was originally the case. But the statistical office does not have the information to check this.[59)]

An example[60)] of this situation is found in Tables 8.1 and 8.2: we find that cell phone id_1 is active in all three hour blocks h, $h+1$ and $h+2$. The location at $h+1$ of this cell phone is missing. But in the approach we described above we impute its goal cell when we consider the hour block pair h and $h+1$, and its start cell when we deal with the hour block pair $h+1$ and $h+2$. These imputations are made independently of each other and therefore may result in different cells. But because we deal with a single cell phone there should actually be a single location in hour block $h+1$. In our approach such 'mistakes' are unavoidable.

59) The telecom provider would be in the position to do this. But we assume that this party is not involved in the process described here.

60) In fact the only one!

If the imputations described above are repeated several times – in a bootstrap procedure – it yields independent estimates of the densities f_C^h and Markov matrices M_C^h, and hence gives an idea of the sensitivity of the estimates. Of course, these estimates can be averaged to imply yet another estimate of these quantites, one with a higher precision.

E Adjustment problem for Markov matrices and cell densities

Here we discuss a problem with the identities (37), for $h = 1, \dots, 23$, the number of consecutive hour blocks in the observation window of a full day, consisting of 24 hour blocks. The problem is due to the fact that the densities have been measured in one way (using total presence derived from cell phone activity) whereas the directions in which the cell phones (and their users) are moving is conditional on where they are. Therefore the densities and the Markov matrices can be considered as independent variables, so that it is most likely that (37) does generally not hold exactly, only approximately. In fact, this is an interesting situation, as this phenomenon will be observed even if there are no measurement errors. This is not what typically happens in practice when constraints are violated.

We start combining the equalities (37), for $h = 1, \dots, 23$, into a single matrix equation:

$$\begin{pmatrix} f_C^2 \\ f_C^3 \\ \vdots \\ f_C^{24} \end{pmatrix} = \begin{pmatrix} M^1 & 0 & \cdots & 0 \\ 0 & M^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M^{23} \end{pmatrix}' \begin{pmatrix} f_C^1 \\ f_C^2 \\ \vdots \\ f_C^{23} \end{pmatrix} \tag{E.1}$$

As a shorthand for (E.1) we introduce

$$\mathsf{f}_2 = \mathfrak{M}' \mathsf{f}_1 \tag{E.2}$$

where

$$\begin{aligned} \mathsf{f}_1 &= (f_C^1, \dots, f_C^{23})', \\ \mathsf{f}_2 &= (f_C^2, \dots, f_C^{24})', \end{aligned} \tag{E.3}$$

and $\mathfrak{M}$ is the untransposed matrix in (E.2). Note that $\mathfrak{M}$ is a Markov matrix, so that holds

$$\begin{aligned} \mathfrak{M} &\geq 0 \\ \mathfrak{M}\, \iota_{23n} &= \iota_{23n}, \end{aligned} \tag{E.4}$$

where 24 is the number of hour blocks in the observation period W, n is the number of cells and ι_k is the all ones column vector of length k (see (8)).

In fact (E.1) does represent the ideal situation. In practice it is likely that

$$\hat{\mathsf{f}}_2 \neq \hat{\mathfrak{M}} \hat{\mathsf{f}}_1, \tag{E.5}$$

where $\hat{\mathfrak{f}}_1$, $\hat{\mathfrak{f}}_2$ and $\hat{\mathfrak{M}}$ are 'measured' quantities.

The idea to remedy this is to change $\hat{\mathfrak{f}}_1$, $\hat{\mathfrak{f}}_2$ and $\hat{\mathfrak{M}}$ iteratively in such a way that (E.2) holds, with as little modifcation of the original values as possible. Of course, this only has a meaning if metrics are introduced to measure the distance of vectors and matrices. We assume Euclidean metrics $d_k(\cdot,\cdot)$ for the vectors and matrices (with appropriate k), generalizing $d_2(\cdot,\cdot)$ in (15).

Because $\hat{\mathfrak{f}}_1$ and $\hat{\mathfrak{f}}_2$ have a considerable overlap, we rather look at the vector $\mathfrak{f} = (f_C^1, \dots, f_C^{24})$ and proxies such as $\hat{\mathfrak{f}} = (\hat{f}_C^1, \dots, \hat{f}_C^{24})$ of measured values.

The iteration alternatively modifies the vector $\hat{\mathfrak{f}}$ (conditional on the current value of $\hat{\mathfrak{M}}$) and the matrix $\hat{\mathfrak{M}}$ (conditional on the current value of $\hat{\mathfrak{f}}$. The iteration starts by fixing $\hat{\mathfrak{M}}^{(0)} = \hat{\mathfrak{M}}$ and then computing a modification $f^{(1)}$ of $f^{(0)} = \hat{\mathfrak{f}}$ in such a way that holds

$$f_2^{(1)} = \left(\mathfrak{M}^{(0)}\right)' f_1^{(1)}, \tag{E.6}$$

where $f_1^{(1)}$ and $f_2^{(1)}$ are vectors assembled from $f^{(1)}$, in such a way that $f^{(1)}$ is as close to $\hat{\mathfrak{f}}$ as possible.

In the next step we look for a matrix $\mathfrak{M}^{(1)}$ satisfying

$$f_2^{(0)} = \left(\mathfrak{M}^{(1)}\right)' f_1^{(0)}, \tag{E.7}$$

which is as closely to $\hat{\mathfrak{M}}$ as possible, where f_1 and f_2 are the original observations, that are fixed. So we now have an updated matrix for $\mathfrak{M}$, namely $\mathfrak{M}^{(1)}$.

We now repeat this by taking similar steps, starting with the updated constraint of the type (E.6), namely

$$f_2^{(2)} = \left(\mathfrak{M}^{(1)}\right)' f_1^{(2)}, \tag{E.8}$$

where the matrix $\mathfrak{M}^0$ has been updated to $\mathfrak{M}^{(1)}$ and the vector $f^{(1)}$ is updated to $f^{(2)}$. Then this vector is used to obtain an updated version $\mathfrak{M}^{(2)}$ of $\mathfrak{M}^{(1)}$ using the equivalent of (E.7). Etc. In this way we produce two sequences: $f^{(0)}$, $f^{(1)}$, $f^{(2)}$, $f^{(3)}$, … and $\mathfrak{M}^{(0)}$, $\mathfrak{M}^{(1)}$, $\mathfrak{M}^{(2)}$, $\mathfrak{M}^{(3)}$, …. Small change of successive values of the vectors and matrices (less than previously specified thresholds δ_f and $\delta_{\mathfrak{M}}$) can be used as a stopping criterion, assuming convergence of the iteration. In [7] a similar procedure is presented, however in a totally different context. In this paper it is claimed that the iterations converge. We have not checked that the procedure described here indeed converges (possibly requiring additional conditions). We leave this issue to be sorted for future research.

F Sensitivity functions and sensitivity areas for cells

In the approach in the present paper we did not assume anything about the sensitivity of each cell c in the network. As we did not have any telecom data at our disposal, it would be strange to build a theory using these concepts. The discussion would then have been rather abstract. Without assuming anything about these cell sensitivities[61] we were able to get some geographical results derived from the telecom data, more specifically the total presence per hour block. By using geometric interpolation we came across Voronoi polygons for each of the cells in the network. In hindsight it may seem as if Voronoi polygons are special kinds of sensitivity areas, but they are not.

The approach taken in the present paper to distribute total presence (see e.g. Section 2.5) per cell per hour block over areas using geometric extrapolation (and hence using Voronoi polygons) has the advantage that it is possible to estimate dynamic population densities, however crude. With cell sensitivity information available these initial estimates can be improved. In the present appendix we want to indicate how this could be done by modifying the approach in this paper somewhat.

The present appendix is intended to indicate how the approach in the present paper can be enhanced in case information is available about the sensitivity of cells c, either in the form of sensitivity functions ζ_c or sensitivity areas Z_c. A sensitivity function ζ_c for cell c provides information about the strength of a signal emitted by c at a location x near c. We assume that ζ_c is normed, in the sense that

$$\int_L \zeta_c(x)\, d\mu(x) = 1, \tag{F.1}$$

for each cell c, where μ is a measure on L. We can view $\zeta_c(x)$ as a conditional density:

$$\zeta_c(x) = f(x|c). \tag{F.2}$$

In practice we are dealing with a reverse conditional probability density $f(c|x,t)$ which can be expressed in terms of the $\zeta_{c'}(x)$ and $f^t(x)$, the dynamic population density at location $x \in L$ at time t, that we wish to estimate:

[61] One can claim that tacitly assumptions have been made about cell sensitivities, of course, for instance that each cell is omnidirectional and that sensitivity areas have been clipped so as to not overlap. But this is unfair as the approach taken was by using straight geometric extrapolation, and not bother about sensitivity areas. If Voronoi polygons are viewed as a kind of sensitivity areas this is the result of an (inadequate) interpretation afterwards. And in its defence one can ask: what else could have been done without knowledge about cell sensitivities? It is comparable to situations where doing nothing actually means doing something, relevant for the problem at hand. As an example, consider the decision to ignore missing values in a data set (that is, not to impute them) when using these data for making estimates. This decision obviously influences the estimates obtained.

$$f(c|x,t) = \frac{\zeta_c(x) f^t(x)}{\sum_{c'} \zeta_{c'}(x) f^t(x)}. \tag{F.3}$$

The sum in the denominator of (F.3) is over the cells c' for which $\zeta_{c'}(x) \geq \theta$, where θ is a threshold, to make sure that the signal emitted by cell c is strong enough to be picked up by a cell phone. We come back to this topic below, when discussing sensitivity areas.

In practice we are also dealing with discrete time and instead of 'global'[62] densities like f^t in (F.3) we have local densities f_c^h per hour block h and per cell c. We have for the conditional density equivalent to $f(c|x,t)$ as defined in (F.3) for hour block h:

$$f(c|x,h) = \frac{\zeta_c(x) f_c^h}{\sum_{c'} \zeta_{c'}(x) f_{c'}^h}, \tag{F.4}$$

where f_c^h is the cell density at cell c in hour block h as defined in (5).

A sensitivity area of a cell c is the set of locations where a cell phone would be able to connect with c. A sensitivity area Z_c^θ for c can be defined when we have a sensitivity function ζ_c: it is the set of those locations x for which $\zeta_c(x) \geq \theta$, for some threshold $\theta > 0$. The threshold θ is taken in such a way that signals $\geq \theta$ emitted from c guarantee good reception, whereas signals $< \theta$ do not. So we define the sensitivity area for c as $Z_c^\theta = \{x|\zeta_c(x) \geq \theta\}$.

Instead of working with sensitivity functions ζ_c we can also work with sensitivity areas Z_c^θ. We then arrive at an approach that is close to the one presented in the present paper. The mass (= total presence) f_c^h per cell c for hour block h is then distributed uniformly over the sensitivity area Z_c^θ instead of the Voronoi polygon V_c. The sensitivity areas are not likely to form a partition of the country L, unlike the Voronoi polygons. The Z_c^θ ideally cover L. But in practice they may not cover all of L. The bits not covered are somewhat isolated areas, hardly visited by people, if visited at all.[63] We shall say that $\{Z_c^\theta | c \in \mathcal{C}\}$ is a near cover of L. This means that working with sensitivity areas is somewhat different from working with Voronoi polygons, which we shall illustrate.

Once the total presence is distributed over the sensitivity areas we have to compute the distribution of the entire 'mass' across L in case we want to produce a graphical density representation for hour block h. This means that the overlap of sensitivity areas have to be taken into account. In such overlap areas the mass in each of the overlapping parts of the sensitivity areas involved have to be added up. For the parts of L that are not covered by sensitivity areas the density is unknown. The total density thus obtained can be smoothed, as in Section 7.3 for Voronoi densities and represented as a heatmap. So both approaches are essentially the same, differing only in nonessential details.

This is also the case when 'density mass' f_c^h distributed over density areas Z_c^θ is to be 'donated' to statistically meaningful areas as municipalities. This happens in essentially the same way as in

62) That is, defined for L, or a major portion of it, not locally, for cells and their immediate vicinities.
63) Sometimes such areas are referred to as cell phone dead spots or areas with no cell service.

case of Voronoi polygons (as discussed in Section 7.2). The main difference is that the sensitivity areas do not ncessarily form a partition, which is not essential. However the idea is that the density mass of such an area, say Z_c^θ, is donated by a municipality M_j proportional to the size $|D_c \cap M_j|$ of the overlap $D_c \cap M_j$. See Section 7, in particular Sections 7.1 and 7.2, for a discussion of the corresponding Voronoi case.

The equivalent of (F.4) for areas W which are parts of the refinement of the sets of the (near) cover $\{Z_c^\theta\}$ is the following probability

$$f(W|c,h) = \frac{f_c^h}{\sum_{c'} f_{c'}^h}, \tag{F.5}$$

where the sum in the denominator is over those cells c' with $f_{c'}^\theta > \theta$.[64] The signal sensitivity is implicitly used through the sensitivity areas themselves.

The part allocated to c is proportional to its total presence f_c^h. The simplifying assumption used is that all points within a sensitivity area receive sufficiently strong signals, whereas the signals for locations outside this area are too weak for a reliable link between c and a cell phone.

As in case of Voronoi polygons, we have similar problems with sensitivity areas near the border of L in the sense of clipping these areas for landbased cells that partly cover a water mass such as a big lake or a sea. Contrary to Voronoi polygons near the border of the country there is no danger that sensitivity areas can get too big. This is because the sensitivity areas are defined independently of each other. This is different from Voronoi polygons where adjacent specimens do depend on each other as they share a common border.

Hopefully this brief description about how to modify the approach presented in the present paper is sufficiently clear so that it can be applied in case sensitivity data about cells become available, in combination with telephone data about the use of cell phones. Then it would be possible to compare the estimates based on various models and approaches.

[64] See Section 7.2 for a discussion of refinements of two partitions. In this case we are dealing with a (near) cover consisting of sets that may overlap. Through computing a refinement the union of the sensitivity areas $\{Z_c^\theta\}$ is subdivided into nonoverlapping parts. The locations in such a part are similar with respect to signal strengths of the various cells: above or below the threshold θ.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information