Dynamic Location of Phone Call Clusters

Problem presented by

José Gil

Motorola

Executive Summary

When mobile handsets are making a call, a measurement report is sent to the serving base station periodically which includes the signal strengths to the base station and the next six strongest signals of the surrounding base stations. Motorola asked the Study Group if it was possible to say whether we could use this information to infer if phone calls occur in clusters and if it was possible to determine the locations, size and other features of these clusters. The Study Group found clusters in 'signal space,' that is, handsets reporting similar signal strengths with the same base stations and explored methods of locating these clusters geographically.

Version 1.0
05 June 2008

Report coordinator

Tristram Armour (Industrial Maths KTN)

Contributors

David Hewett (University of Oxford)
Peter Hicks (University of East Anglia)
Peter Howell (University of Oxford)
Juliusz Fiedler (University of Warsaw)
John Ockendon (University of Oxford)
Ricardo Pachon (University of Oxford)
Honora Smith (University of Southampton)
Eddie Wilson (University of Bristol)

ESGI64 was jointly organised by
Heriot-Watt University
The Knowledge Transfer Network for Industrial Mathematics
The International Centre for Mathematical Sciences

and was supported by
Engineering and Physical Sciences Research Council
The London Mathematical Society
The Institute of Mathematics and its Applications
The European Journal of Applied Mathematics

Contents

1 Introduction

A common problem in cellular systems is identifying the locations of users on the system. Although mobile handsets incorporating GPS systems do exist, their use is far from widespread and likely to remain so until battery-life problems are solved. In addition, GPS does not function well inside buildings or in heavily built-up areas, and may only provide a crude estimate of location.

Position estimation is often performed to provide individual users with location-specific information - for example proximity to shops or stations - or to provide contextual advertising or mapping assistance. The Study Group was asked to analyse the traffic distributions and densities rather than the locations of individual subscribers: is the traffic evenly spread over the serving area, or are there localised clusters of heavy traffic for example, at station or theatre exits? Are these clusters static, or do they change over time? What is the size of these clusters, and how accurate are the estimates of the cluster location and size? Furthermore, can the distribution of the subscribers be classified as in-building or outdoor by observing the data? Additional classification of the traffic would be to cluster the users in terms of their mobility (static/pedestrian/vehicular) and distribution in the vertical.

The information gathered from the clustering analysis would be invaluable for network operators wishing to determine where they should be integrating additional network capacity, for example through the introduction of picocells, femtocells, and WLAN access points. Combining the cluster information with call models and sample tariffs can provide detailed business plans to support analysis of likely return on investment.

Mobile handsets (MS) are usually in contact with one or more base stations (BS) during and between calls. The mobile measures the received signal strength from nearby base stations and attempts to access the BS with the strongest signal when a call is to be established. As the user moves around the system, the varying signal strengths received from neighbouring BSs are recorded by the mobile and reported to the serving BS. If the user moves out of the coverage area of the serving BS, a handover can be performed which allocates the MS to a new serving BS.

The MS sends the RSSI (Received Signal Strength Indication) information back to the serving BS in the form of periodic measurement reports (MR). The primary function of the MRs is for handover and mobility control; however it is possible to sample and store the MRs by analysing the communication links from the BS to its controller (RNC), or by call-trace techniques at the BS itself. In this way a large number of MRs can be captured from the entire population of MS in an area being served by a group of cells. Many techniques have been proposed and implemented which attempt to derive the MS location from the MRs, commonly based around a triangulation approach or TDOA (time difference of arrival). Although these techniques can be effective in some situations, there are a number of problems to be solved. Firstly, there may only be RSSI information from the serving cell (and no other neighbours) for many MS. Secondly, the BS in a GSM system are unsynchronised and their timing references may drift relative to each other, introducing

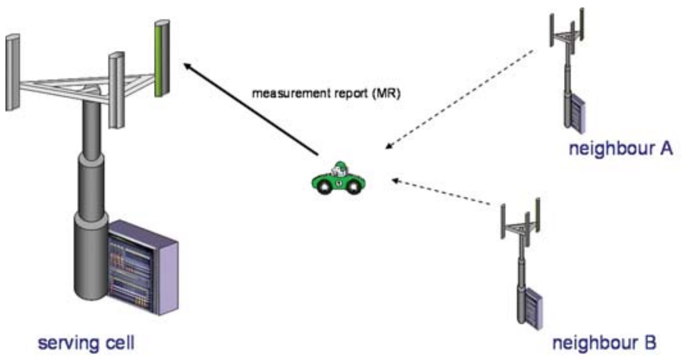

Figure 1: A measurement report being sent to the serving base station

errors in the timing measurements reported by the MS. Thirdly, the signal received by the MS is attenuated by other factors such as Rayleigh and Rician fading caused by obstructions, multipath scattering, and also Doppler effects caused by mobility. The combined effects of all these factors need to be considered when producing a location estimate for a specific MS.

1.1 Data

(1.1.1) The GSM data we were given was a subset of MRs reported to a base station in a ten hour period. Every 480 ms, each handset making a call sends an MR to the serving BS containing the signal strength (power) of the serving BS and 6 strongest signals from the neighbouring BSs, the signal quality (related to the bit error rate) and the timing advance (TA) to the serving base station. The timing advance is an integer and is a synchronisation variable which enables the handset to send the data to the correct time slot allocated for the handset at the serving base station. If $TA = n$, the handset is approximately between $550n$ and $550(n+1)$ metres from the serving BS along the strongest signal path. (Each TA unit represents 3.69 μs shift in the time slot which corresponds to approximately 1100 m difference to the round-trip distance.) We would like to use this information together with the signal strengths to produce a better estimate of the location of the handsets.

(1.1.2) There are two types of base stations, some are omni-directional *i.e.* the power output is evenly spread over 360 degrees from the BS whilst others are directional. The directional antennae concentrate their signal in a 120 degree arc. So at any one BS location, we could have one omni-directional BS or three directional BSs. Together with the MR data, we had the location (latitude and longitude coordinates) and power output of each BS and the azimuth for the directional antennae BSs.

1.2 Signal space and physical space

(1.2.1) Although each MR only reported up to six of the strongest signals from the neighbouring BSs, there were 52 different neighbouring BSs recorded in the whole data. We can think of a measurement report as a projection from the 'signal space' in $\mathbb{R}^{53}$ to a seven-dimensional subspace. Each point in this signal space may map to several different points in the 'physical space' where the handsets are located.

1.3 Proposed solution

(1.3.1) Using a high dimensional cluster analysis technique we can try and identify clusters in the signal space. Using extra information in the MRs, we can start to analyse the cluster and possibly split the cluster up further depending on what we find. For example, if the cluster moves rapidly in signal space with time, we could assume that the handsets are moving. However, this may not be true in all cases – for example, if a rumour spread through a crowd, a cluster in signal space would be moving even though the people themselves remain still. If we find that some handsets of the cluster have a weak signal quality (related to the bit error rate) compared to the others, we could conclude that there is a cluster indoors and one outdoors.

(1.3.2) Further work is required to verify whether clusters in signal space correspond to clusters in physical space. However, with some careful analysis of the clusters and their MRs we could map these signals into physical space using for example, the triangulation method proposed in this report in section 2.3.

2 Details

2.1 Propagation models for the base station signals

(2.1.1) In free space we can relate the signal power to the distance from the BS by

$$P_R = G_T G_R \left(\frac{c}{4\pi r f} \right)^2 P_T. \tag{1}$$

where G_T is the gain in the transmitter, G_R is the gain of the receiver, f is the frequency of the signal, c is the speed of light, r is the distance between the transmitter and the receiver and P_T is the power of the transmitter.

(2.1.2) In the presence of a flat plate providing a ground reflection, destructive interference implies

$$P_R = \left(\frac{h_T h_R}{r^2} \right)^2 P_T. \tag{2}$$

where h_T is the height of the transmitter and h_R is the height of the receiver, both assumed small compared to r. In reality buildings and other clutter scatter the signal and the radiation is not uniform in all directions.

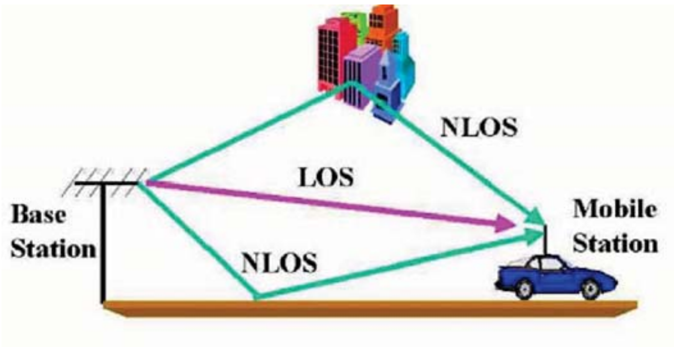

Figure 2: A signal from a base station can travel either by a line-of-sight path (LOS) or by non-line-of-sight (NLOS) paths

(2.1.3) Using the MR data, we can plot a real picture of the relation between the signal strength of the serving BS and the distance from it using the timing advance variable. As we can see in Figure 3, the data may be noisy and we therefore we wouldn't be able to do any trend fits which could be useful for developing propagation models. However, there is some observable structure in the data, namely the stripe at $TA = 30$, and the clusters centred about (TA,Signal Strength) $= (40, 17)$ and $(50, 16)$.

2.2 Constraint satisfaction technique

(2.2.1) Let us first assume that the signal strength decays monotonically with distance from the BS and that this decay is the same in each direction, and the same for each BS. With this assumption and the sequence of strongest signal strengths from the surrounding BSs, we can geometrically constrain the region in which the MS is located. Consider the idealized case of a regular hexagonal cell.

(2.2.2) To determine the possible region in which the MS is located, we construct the perpendicular bisectors of each pair of base stations. For example, if the signal strength from BS_1 is stronger than BS_2, then we can conclude that the handset must lie on the side of the perpendicular bisector closer to BS_1. We repeat with all pairwise combinations available for that MR. For a sufficiently large cell, we can use this information together with the timing advance and find a smaller region where the handset must be (see Figure 5). In the regular hexagonal case, the third strongest BS constrains the handset to one of twelve sectors of the serving cell. However, due to

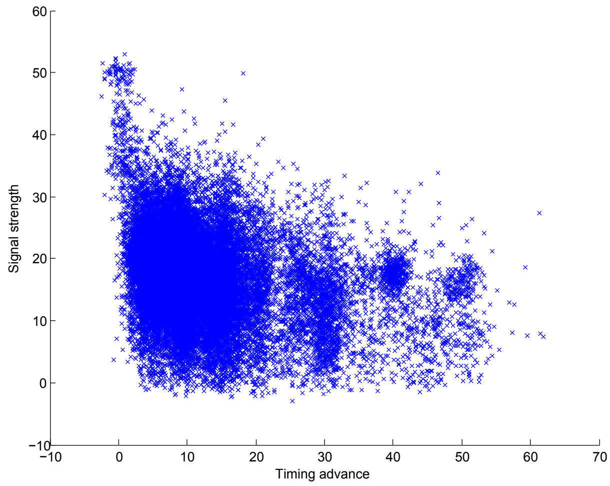

Figure 3: Signal strengths of the serving BS plotted against the timing advance

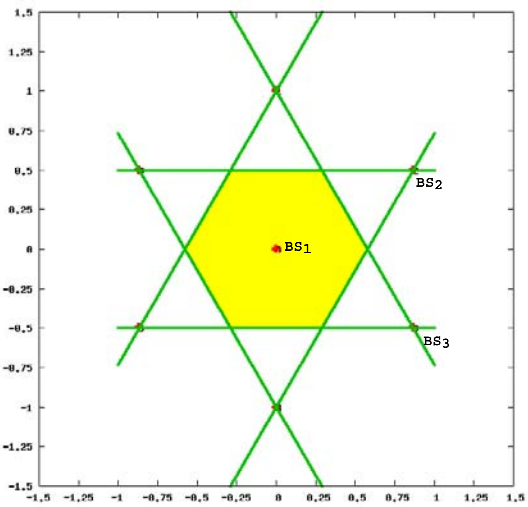

Figure 4: Bisecting between the serving BS and the neighbouring BSs

the symmetries of of a regular hexagon, considering additional indices provides no additional information.

(2.2.3) With an irregular distribution of the neighbouring cells, we break the the hexagonal symmetries in the BS distribution and increase the number of possible regions (Figure 6). In this case, it would be best to start simple and look at the three BSs with the strongest signal (as in Figure 7).

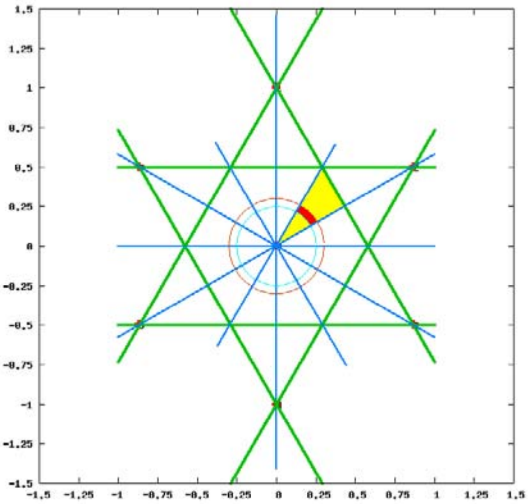

Figure 5: Locating a handset using constraint satisfaction and TA

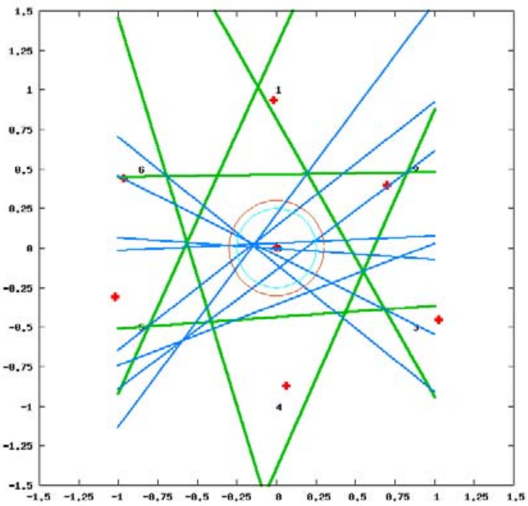

Figure 6: Irregular distribution of neighbouring BS

(2.2.4) It is likely that in an inner city environment the assumption of monotonic signal strength decay does not hold because the signal strength of a BS depends more on the signal path than the absolute distance from that BS. There were many reports in the serving cell which reported neighbouring BSs that were not one of the six closest to the serving BS. This means that we overconstrain the region and we must somehow choose which constraints are most important and try and minimise the number of soft constraints that need to be broken to solve for the region. This is unlikely to produce any useful result in a city environment but in a flat rural environment, given a set of MR data we could count the number within

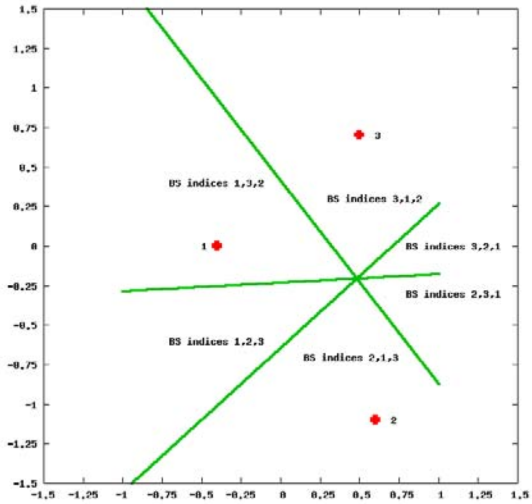

Figure 7: Using three BSs with the strongest signal

each sector and look for clusters using this method.

2.3 Least squares estimate for handset location

(2.3.1) In order to locate a handset with greater precision than the above method, we need to consider the actual values of the reported signal strengths and not just the order of the signal strengths to each neighbouring BS.

(2.3.2) Assuming an obstacle-free radial propagation model, the isopower curves around the i-th BS for a constant receiver height h_R are circles with radius

$$r_i = \left(\frac{h_{T_i}^2 h_R^2 P_{T_i}}{P_{R_i}} \right)^{1/4} \tag{3}$$

where h_{T_i} is the transmitter height of the i-th BS, P_{T_i} is the transmission power of the i-th BS and P_{R_i} is the power of the received signal from the i-th BS.

(2.3.3) If we knew the handset heights, h_R, the handset can then be located exactly by triangulation between multiple BS's and in the case with no clutter and no noise, we can find the mutual intersection point of all of the circles (Figure 8).

(2.3.4) However with real data, the computed r_i may no longer give a point of mutual intersection due to noise and clutter in the environment. One way to proceed in this case would be to obtain an estimate for the handset location using a least-squares minimisation approach. We used the cost function

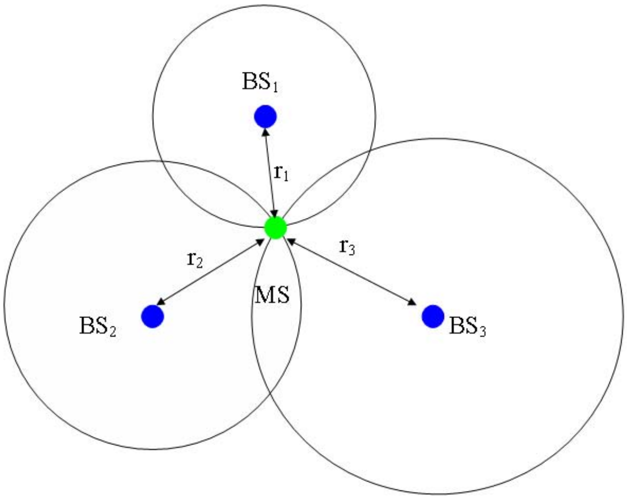

Figure 8: Triangulating the position of the handset

$$f(x, y) = \sum_{i=1}^{n} \left(\sqrt{(x - x_i)^2 + (y - y_i)^2} - r_i \right)^2 \tag{4}$$

where n is the number of base stations involved in the data, x_i and y_i are known (x, y) positions of the base stations and r_i is the distance estimated by applying equation 3. *i.e.* The cost function minimises the sum of squared perpendicular distances from the idealised circles. h_R will have to be estimated and the Study Group assumed a value of 1.5 m after consultation with José Gil.

(2.3.5) To test this method, we generated sample data by applying a random relative error (of up to ±50%) to the expected power levels. The results of this can be seen in Figure 9.

(2.3.6) Note that in the obstacle-free case we need only 3 BSs to determine the handset location exactly. We have data from 7 BSs so the system is overdetermined. The Study Group investigated briefly whether we could use this extra information to determine unknown characteristics of the environment such as the height of the transmitters at the base stations. Extending this idea, it was proposed that we could use measurement reports to perform tomography on the domain. *i.e.*Find the isopower curves in the domain. However, further research will be needed to determine whether this is possible. To illustrate this point, we ran the triangulation method with the transmitter heights of the base stations also as unknowns in the system of equations. We managed to recover the positions and the transmitter heights (See Figure 10).

(2.3.7) We ran a test of the triangulation method on real data using a sample

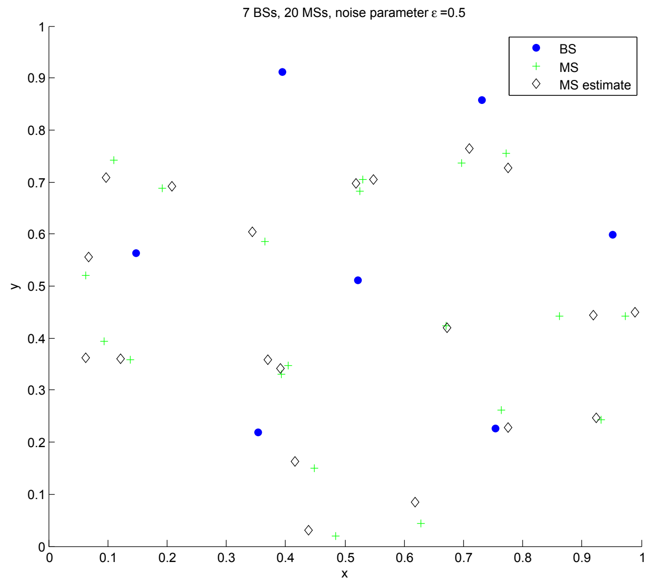

Figure 9: Triangulating positions of handsets with known transmission heights

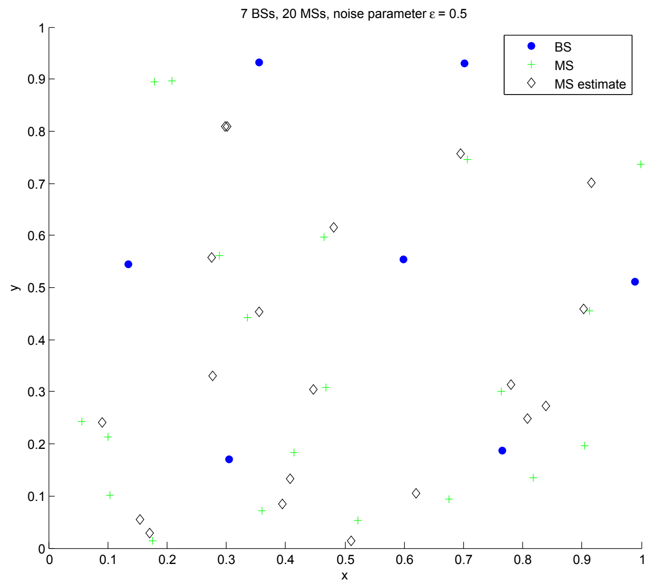

Figure 10: Triangulating positions of handsets with unknown transmission heights

of 200 MRs which included 21 different base stations. The results can be seen in Figure 11. Some clustering can be seen in what appears to be built up area but further research and information is required to validate this.

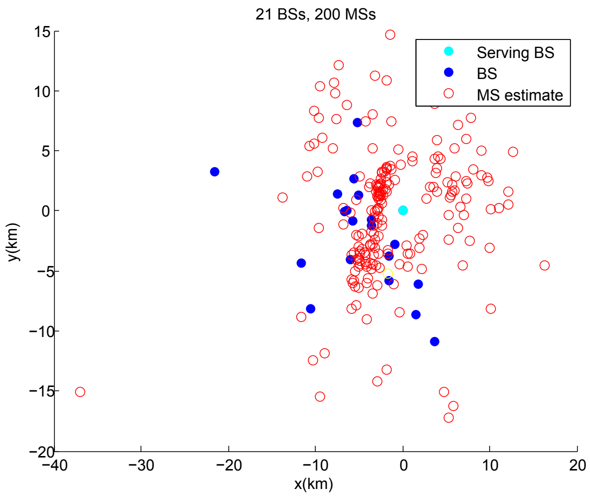

Figure 11: Locating handsets using real data

(2.3.8) If fading produces error of some additive factor in the signal strength, it would be worth modifying the cost function in Equation 4 to involve the square difference in logarithms instead.

(2.3.9) A final remark about this method is that the cost function used at the Study Group is not convex and is likely to lead to multiple minima. Of course the problem may not be convex in which case we should not go for an artificially convex cost function that eliminates the other minima. However, we should not introduce any unnecessary non-convexity because that will tend to make the numerical minimisation more difficult.

2.4 Cluster analysis in signal space

(2.4.1) Ideally, to determine whether phone call clusters exist, we would use co-ordinates of the mobile handsets. As such data is not available at present, we have to consider an alternative. The data that is available includes signal strengths from neighbouring base stations, and also timing advance and signal quality corresponding to the serving BS. It is considered that handsets close by would have the same readings with such variables, unless there is a high degree of distortion of the signals by buildings. We therefore consider finding clusters in signal space, rather than physical space, to make the most of the data available.

(2.4.2) The location of clusters is a standard branch of data mining. There are a lot of methods suitable for all kinds of data, such as partitional, hierarchical and density-based clustering with different definitions of a *cluster*. Most of our successful cluster analysis was with *DBSCAN*, a density-based clustering algorithm that looks for static clusters. A detailed description can be found in the Appendix. The Study Group also tried a hierarchical clustering algorithm on Minitab using the whole signal data set.

(2.4.3) To carry out cluster analysis, parameters concerning the nature of clusters have to be specified. How close together items must be to form a cluster must be defined: perhaps within 100 m in physical space, or 1 dB in terms of signal, although this has different implications when close to a base station or further away, because of decay. In a hierarchal clustering algorithm, linkage between items must also be decided upon: do you count average distance between items, distance to the centroid, or Ward linkage, a hybrid measure?

(2.4.4) Runs of 30000 records were input to Minitab clustering analysis, using Euclidean distance measurements and Ward linkage. No particularly strong clusters were found, even with only 20% similarity (calculated from distance apart d, using the formula $100X(1 - d/d_{max})$ where d_{max} is the maximum distance between a data point a point in a cluster. We can conclude that cluster analysis using the whole data was not very successful. This is because the points are very sparse and clusters are found at the origin as zeros must be used to pad for missing data.

(2.4.5) There was more success when we looked for clusters in the two-dimensional subspace of signal space formed by comparison of pairs of base stations. Each plot in Figure 12 shows the intensity of signals for different BS pairs. Each point represents the signal strength of a certain handset to the two stations. Applying DBSCAN on each dataset, we obtained clusters coloured with blue, green, red, yellow or purple. Black points are tagged as noise and are not classified in any cluster.

(2.4.6) Ultimately, we want to look for clusters in the whole signal space, which corresponds to clustering in a high dimensional space. See references (5.1.5), (5.1.6) for more information on high-dimensional clustering. CLUTO, a graph-partitioning algorithm, would be a good algorithm to start with (see reference (5.1.7)).

(2.4.7) Finally, we need to map clusters from the signal space to the physical space. A method would be to map the signals in a large group of homogeneous samples (from cluster analysis) using the triangulation method in Section 2.3 with an appropriate propagation model for the serving base station. Ideally, instead of locating each handset individually as with the original algorithm, we want to apply it on the cluster itself where the output would be an area in physical space which represents the location of the cluster. This extended model could be validated using real signal

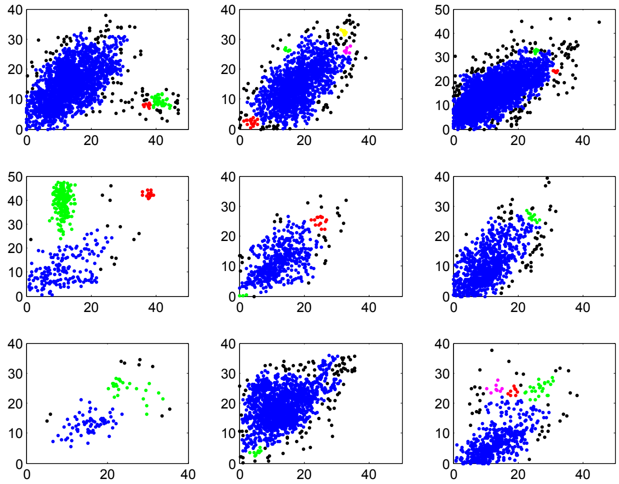

Figure 12: Cluster analysis using DBSCAN on pairs of commonly reported neighbours

strength data and location data from experiments around the base station. However, these experiments should be repeated from time to time because propagation path losses will change due to changes in vehicular traffic and cityscape.

3 Conclusion

3.1 Final remarks

(3.1.1) The two approaches we suggested were to

(a) Locate individual handsets in physical space and look for clusters in the physical space.

(b) Identify clusters in the signal space data and map these to the physical space.

(3.1.2) Cluster analysis on the MR data can tell us if clusters occur in signal space. A detailed analysis of the data is required in order to conclude that the cluster in signal space is indeed a cluster in the physical space. In one particular experiment in (5.1.2) there have been measured signal strength differences of up to 50 dB in a distance of 30 m in an inner city area. However, there are many base stations in a built up area and it is not unrealistic to suggest that there will be at least one base station in line-of-sight with the clustered handsets.

(3.1.3) Cluster analysis cannot be used to find clusters in real time because the number of dimensions is high even for a small set of MR data. However, the triangulation method is fast enough to be run in realtime for a reasonably sized data set. However it is unclear at present how accurate this method is at locating clusters.

(3.1.4) One possible method to find clusters in real time would be to run hypothesis testing on the arrival of calls (assuming Poisson arrivals with rate λ where λ is known). By observing the number of handsets connected to the BS in intervals, we can test with an alternative hypothesis that λ has increased. See reference (5.1.8) for more details on testing the arrival rates of calls.

3.2 Proposed further research

(3.2.1) Using simple propagation models with power proportional to $1/r^2$ or $1/r^4$ will not produce satisfactory results in built up areas. To produce better propagation models we recommend Motorola gather experimental data of signal strengths near one or more base stations. Using this information we could also validate the triangulation method results in Figure 11.

(3.2.2) The success of a clustering algorithm depends greatly on the data set. We recommend that Motorola investigate high-dimensional techniques perhaps starting with graph partitioning algorithms on the MR data. Another type of clustering algorithm worth investigating is an overlapping clustering algorithm such as *Fuzzy c-means* which allows data points to be a member of one or more clusters.

(3.2.3) Using extra information in the MRs, we could analyse the clusters in signal space and possibly split the cluster up further depending on what we find. For example, if the cluster moves rapidly in signal space with time, we can assume that the handsets are moving. If we find that some handsets of the cluster have a weak signal quality (related to the bit error rate) compared to the others, we could conclude that there is a cluster indoors and one outdoors. We recommend that once Motorola have found clusters in signal space, they investigate the extra MR data available for the clustered handsets.

(3.2.4) Unfortunately we did not have time to explore mappings of the clusters in signal space to the physical space during the Study Group and further research will be needed to assess approaches.

4 Appendix

4.1 Conversion of latitude and longitude to $x-y$ plane

(4.1.1) The location of the BSs in the data set were given in latitude and longitude coordinates. As the cost function (Equation 4) is defined on the $x-y$ plane,

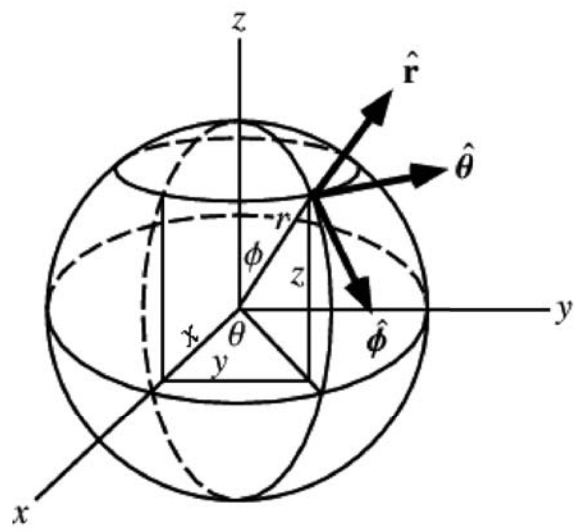

Figure 13: Spherical Coordinates [*Wolfram MathWorld*]

in order to use our least squares method in section 2.3, these coordinates must be projected on to the $x - y$ plane. The conversion from spherical coordinates (r, θ, ϕ) to cartesian coordinates (x, y, z) is

$$\begin{aligned} x = & \quad r \sin\phi \cos\theta \\ y = & \quad r \sin\phi \sin\theta \\ z = & \quad r \cos\phi \end{aligned} \tag{5}$$

where r is the radius of the Earth, θ is the longitude and $\pi/2 - \phi$ is the latitude of the BS (Figure 13). Alternatively, we can recast the problem in spherical coordinates but we did not pursue this.

(4.1.2) To minimise the distortion caused by the projection in the next step, we rotate the sphere so that the serving BS (with latitude $\pi/2 - \phi_0$ and longitude θ_0) lies on the South Pole.

(4.1.3) The (x, y, z) coordinates were now mapped onto the $x-y$ plane (by stereographic projection) about the North Pole with the following formula

$$(X, Y) = \left(\frac{x/r}{1 - z/r}, \frac{y/r}{1 - z/r} \right) \tag{6}$$

and $(2rY, 2rX)$ are the scaled projected positions with the y-axis pointing due North. The scaling is an approximation as the scaling depends on the distance from the origin but it should perform well for a small area of interest.

4.2 Density-based clustering algorithms

(4.2.1) The key idea in a density-based clustering algorithm is that for each core-point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points (N_{min}). The shape of a neighborhood is determined by the choice of a distance function for two points p and q, denoted by $d(p, q)$. In order to define what we mean by a cluster, we start with some definitions.

(4.2.2) Given d, we define the ϵ-neighbourhood of a point p, denoted by $N_\epsilon(p)$ by

$$N_\epsilon(p) = \{q \in D | d(p, q) \leq \epsilon\} \tag{7}$$

(4.2.3) We say a point p is *directly density reachable* from a point q w.r.t. ϵ, N_{min} if

- $p \in N_\epsilon(q)$ and
- $|N_\epsilon(q)| > N_{min}$ both hold.

In words, p is *directly density reachable* from a point q if there are more than N_{min} points within ϵ of q and p is one of them. Note that this condition is not symmetric.

(4.2.4) A point p is *density reachable* from a point q w.r.t. ϵ and N_{min} if there is a chain of points $p_1, ..., p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i $\forall i < n$. Again, this is not symmetric due to the definition of *directly density reachable*.

(4.2.5) Finally, a point p is *density connected* to a point q w.r.t. ϵ and N_{min} if there is a point r such that both, p and q are density-reachable from r w.r.t. ϵ and N_{min}.

(4.2.6) Given ϵ and N_{min}, we define a cluster C in a density based algorithm as a subset which satisfies a

(a) *Maximality* condition (if $p \in C$ and q is density-reachable from p then $q \in C$)

(b) *Connectivity* condition (for $p, q \in C$ then p and q are connected)

Not all points will be members of a cluster, we define these points as noise. Density-based algorithms are ideal for clustering noisy data because we define a cluster using parameters related to the density of the points rather than say in *K-means* where we specify the number of clusters in the data beforehand which makes the algorithm more sensitive to noise.

5 Bibliography

5.1 References

(5.1.1) Ester M., Kriegel H.-P., Sander J., Xu X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, 1996, pp. 226-231.

A reference for DBSCAN.

(5.1.2) Saunders, Simon R.: Antennas and Propagation for Wireless Communication Systems, Wiley.

Introduces the basic concepts and specific applications of antennas and propagation for wireless systems. Covers terrestrial and satellite radio systems in both mobile and fixed contexts.

(5.1.3) P. Andritsos: Data clustering techniques, Qualifying oral examination paper, University of Toronto, (http://www.disi.unitn.it/ periklis/pubs/depth.pdf) 2002.

A simple and general introduction to clustering algorithms.

(5.1.4) Richard O. Duda, Peter E. Hart, David G. Stork.: Pattern Classification, Wiley-Interscience, 2 Sub edition (October 2000)

Another useful introduction to clustering algorithms in Chapter 10.

(5.1.5) A. Strehl; Relationship-based clustering and cluster ensembles for high-dimensional data mining, PhD thesis, University of Texas at Austin, 2002 (http://www.lans.ece.utexas.edu/ strehl/diss/htdi.html).

This reference has a lot of useful background for the problem of high-dimensional clustering.

(5.1.6) E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher, Clustering in a high-dimensional space using hypergraph models, Tech. report, Department of Computer Science, University of Minnesota, 1999.

A reference for high-dimensional clustering using hypergraph partition.

(5.1.7) Michael Steinbach, George Karypis and Vipin Kumar: Comparison of Document Clustering Techniques, KDD Workshop on Text Mining, 2000.

This paper presents the results of an experimental study of some com-

mon document clustering techniques. There are now many variations of CLUTO, one of the first is studied in this paper.

(5.1.8) Brown, L., and Zhao, L.: A new test for the Poisson distribution, Sankhya: The Indian Journal of Statistics, Series A, 64, 611-625, 2002

Hypothesis testing on Possion arrivals.