# Qualitative Positioning for Pervasive Environments

Ian Anderson and Henk Muller

Department of Computer Science, University of Bristol, U.K. {anderson,henkm}@cs.bris.ac.uk

## ABSTRACT

In this paper we present a strategy and a set of algorithms for developing qualitative positioning services that provide a qualitative location optimised for the environment where they are to be deployed. We argue that for many context-aware applications this may be more appropriate than more common quantitative location systems, where the positioning API may make unrealistic demands on the underlying measurement service, and unrealistic promises to the application. We show how a symbolic location system can be learnt from training data in an unsupervised manner. We present experimental results using 802.11 and GSM signal strength levels and wireless beacon data.

*Keywords*: Location fingerprinting, machine learning.

## 1 INTRODUCTION

In 2000 Bahl et al. [3] first demonstrated that it was possible to determine the position of an object by comparing current signal strength levels with those stored in a radio map of the application environment. This technique known by the term 'location fingerprinting' appealed to the ubiquitous and pervasive research communities largely because it built on existing infrastructures such as 802.11 Wireless Networks (WiFi). Recent work however, has started to focus on the use of GSM instead of WiFi as the underlying measurement service [14], [13]. This change of focus has occurred principally because GSM as a technology is more ubiquitous than 802.11 both in terms of coverage and user accessibility; everyone has a cell phone.

In this paper we report on an investigation into the use of ubiquitous signals present in our everyday lives as a method of inferring location. In particular we focus on positioning mobile devices in scenarios typically considered as '*harsh*' such as open, outdoor environments where measurement variation is typically minimal. From a location fingerprinting perspective, the perfect environment is one where the positional dependent measurement will vary widely at different locations but be constant at the same physical location. Hence areas where there is typically only a minimal fluctuation in measurements such as open outdoor environments offer relatively poor positional granularity.

Given positional dependent data such as GSM and 802.11 signal strength, location fingerprinting can be used to recognise areas. There are two ways to look at this: quantitatively and qualitatively. A quantitative positioning system typically takes a geometric view of space using Euclidean or spherical coordinate systems. In a qualitative (or symbolic) location system space is managed as zones which are of interest to human beings. In terms of location fingerprinting, quantitative models require the radio map to be mapped to continuous coordinates, for example X, Y and Z. This requires samples to be collected next to a ground truth. This enables statistical models to be constructed and if required propagation models to be applied to enable finer positional granularity [4]. Qualitative models differ from this approach in that samples do not need to be collected next to a ground truth. Similarity metrics between adjacent readings can be used to create zones that are separable. We assess the use of GSM and 802.11 signals together with wireless beacon information as a means to infer a qualitative location. In this process we have obtained 85,000 data-points. Each data-point consists of a ground truth GPS location (only for testing), seven cell strength readings and where available, 802.11 signal strength readings. The area covers an urban space of 1.5 by 1.5 km, and was collected over an 8-week period.

The rest of this paper is structured as follows. Section 2 provides a review of related work. Section 3 discusses a qualitative method of managing space and demonstrates the suitability of this approach when the underlying measurement service is GSM or 802.11 signal strength data. Section 4 presents an extensible Bayesian network for fusing cellular and 802.11 signal strength data with wireless beacon information. Section 5 discusses how to assess the performance of the radio map. Section 6 reports on a prototype implementation using data gathered from a metropolitan environment.

## 2 BACKGROUND

Location fingerprinting is a positioning technique that has increased in popularity over the last few years. This is largely because it does not require change to the existing network infrastructure, hence cost is low. Other favourable traits include: user privacy, it operates in environments where the Global Positioning System (GPS) would fail (indoors and in dense urban environments) and the number of wireless beacons available in our cities and towns has increased dramatically over the last few years. For example, in 2005 during a war driving survey it was shown that downtown Seattle has a WiFi access point density of 1200 per sq km [6]. The RADAR system [3] was the first to apply this location technique and achieved a median positional accuracy of 2.94 metres using a network of 802.11 (WiFi) wireless access points. Since then there has been much work trying to improve on these initial results by building statistical models, applying complex RF

This research was funded by the UK Engineering and Physical Science Research Council, Equator Interdisciplinary Research Collaboration EPSRC GR/N15986/01 (http://www.equator.ac.uk)

signal propagation models and other tracking/filtering techniques [7], [9], [10].

Recently, Otsason et al. [14] demonstrated that in an indoor environment, using a wide GSM signal strength fingerprint, it is possible to achieve a median positional accuracy of 5-metres. The wide fingerprint contained the signal strength levels for the 6 strongest cells and up to 29 additional GSM channels. This information was obtained using a GSM modem that exported a richer API than most typical GSM cell phones. Laitinen et al. [11] applied location fingerprinting with GSM networks in an outdoor environment, achieving a positional accuracy of 44-metres. The difference in these results is largely due to the minimal variation in signal strength levels in open environments.

Typically, location fingerprinting systems have always required an exhaustive calibration phase where a radio map of the spatial environment is constructed. Generally, once constructed, the radio map is treated as a static entity and new or previously undetected radio beacons are not added once calibration is complete. This poses performance problems when running systems over extensive time periods. To address this limitation, Letchner et al. [13] developed a hierarchical Bayesian framework that enabled new 802.11 access points to be seamlessly integrated into a model of the spatial environment. This approach, designed with large scale WiFi-based coverage and long-term deployment in mind is possible because the radio map is periodically refreshed. Bayesian networks are particularly useful when aiming to provide long-term and large-scale deployment because they take a probabilistic approach to sensor fusion. Therefore they can handle situations where only incomplete data is available such as missing beacon or signal strength information [2]. Cheng et al. [5] have minimised calibration effort by applying positioning algorithms such as a centroid and a particle filter on minimal as opposed to exhaustive data sets. This reduced the time it took to map an entire city neighbourhood to less than half an hour. LaMarca et al. [12] calibrated radio maps using data gathered from application users. This approach had the advantage that the radio map reflected real usage hence popular areas of the application environment were calibrated exhaustively.

These quantitative approaches provide a position and an associated accuracy error that reflects the limits of the underlying measurements. Our work differs from this in that we look to contain the error within the positioning system and return a qualitative location that reflects the best achievable performance given the available measurements and constraints of the spatial environment.

# 3 QUALITATIVE SPACE

Models of space can typically be classified as either topological (qualitative) or more commonly, as coordinate based (quantitative). Quantitative models generally take a geometric view of space with positional information supplied by location services using Euclidean or spherical coordinate systems. Coordinate tuples are processed by the application and behaviour is updated to reflect the new location information. In contrast, topological or symbolic models manage space in a qualitative manner with positional information mapped to human abstractions of physical places usually in the form of spatial zones. The relationships between zones form a topology often expressed as a graph. Application behaviour varies depending upon the symbolic representation of space (zone) that the user is currently located in. When constructing a symbolic model of the spatial environment developers must define spatial zones within the constraints of the underlying sources of positional information. For example, it is not possible to create zones with a physical coverage area that is finer than the granularity of the data produced by the positioning services.

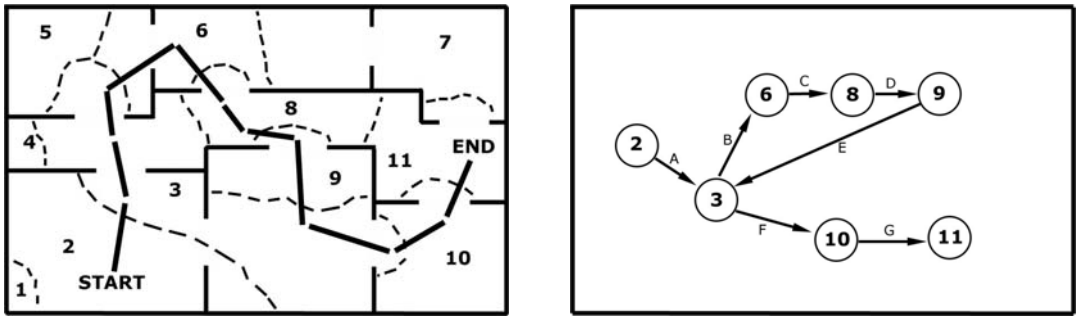
In this section we demonstrate how a qualitative approach to managing space is particularly suitable when the underlying measurement service is GSM or 802.11 signal strength data. This qualitative approach differs from more common quantitative location systems, where the positioning API may, given the available measurements, make unrealistic demands on the measurement device, and unrealistic promises to the application programs.

## 3.1 Logical Management of Space

A spatial environment can be managed in a qualitative manner by introducing the notion of a spatial zone. We use the term '*spatial zone*' to describe a portion of space that, when using a measurement of, for example, signal strength of a wireless beacon, can be distinguished from other areas of space. The area of physical space that a zone symbolises, reflects both the quality of the positional measurements and the spatial environment. Thus zones represent the *finest*, *reliable* position that the measurement service can offer, i.e. if it is possible to reliably determine position within different areas of a zone then the zone should be split into smaller, child zones. Consequently zones do not necessarily cover the same amount of physical space and hence are assumed to be of unequal size.

Our positioning service returns the zone the user is currently located in as their qualitative location. The way that zone membership is determined depends on the type of positional measurements available. For example, Figure 1a shows the layout of an office environment with zones superimposed on it. It also shows the physical path that a user took when walking through this area. In terms of qualitative location, this path simply represents a series of zone transitions in the form of a directed graph as shown in Figure 1b. We use the term '*logical path*' to describe the series of zone transitions equivalent to the physical path.

By constructing logical paths based on users' interactions with the application environment it is possible to infer the relationships between zones. This has the advantage that once sufficient data has been collected it is possible to identify popular paths and invalid zone transitions in an unsupervised manner, making it easier to roll out the system and, over time, improve positioning service performance.




(a) (b)

Figure 1: (a) A typical office floor plan - the spatial environment for a context-aware application. The environment has been partitioned into zones that reflect the performance of the underlying positioning services. (b) The physical path illustrated in (a) can be represented as a directed graph. The nodes in this graph correspond to qualitative locations and the arcs indicate order.

### 3.2 Automatic Zone Creation

We can construct a zone-based representation of a spatial environment in an unsupervised manner. This is a simple offline calibration process. Firstly, the deployer collects samples of positional measurements throughout the application environment. Unlike traditional location fingerprinting calibration, the associated physical positions do not need to be stored with these measurements. Once this training data has been collected it is partitioned into sets of similar measurements. These sets contain the data that will be used to determine zone membership, hence a set, or cluster of training data defines the boundaries of a spatial zone.

In a previous paper [1] we demonstrated how the partitioning of data can be carried out in an unsupervised manner by applying K-means using:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2 \quad (1)$$

where $x_n$ is a vector representing a positional measurement and $\mu_j$ is the centroid of the data points in $S_j$ and $|x_n - \mu_j|^2$ represents the distance between the sample and the cluster centre is used to partition the training data. If, for example, the positional measurements were signal strength levels on a cellular network then $x_n$ would represent a snapshot of these levels for all visible cells. K-means can be initialized with vectors selected at random from the training data. The Euclidean distance for each subsequent sample $x_n$ to the centre of each centroid $\mu_j$ is then calculated. This sample $x_n$ is then added to the centroid that it is closest to. The centroids are then recalculated and the membership of each of the points $S_j$ for each centroid $\mu_j$ is then re-evaluated until there are no further changes in membership. At runtime the qualitative location of a user is determined by finding the cluster most similar to a position dependent measurement taken at the users current, physical location.

This qualitative representation of a spatial environment cannot be created by a simple process of converting an existing quantitative model. That is, mapping a set of Cartesian coordinates to a series of spatial zones. This is because a linear mapping will not be reflective of the limitations of the environment and the positional dependent measurements.

## 4 FUSING POSITIONAL DATA

In our daily lives we are increasingly surrounded by a wealth of information that can be used to infer location. We can use the *sighting* of a particular wireless beacon or combination of wireless beacons to infer information about our current location. For example, if the MAC address of the WiFi access point in your office appears in an 802.11 scan then you can infer that you are *near* your office. By using a combination of currently visible wireless beacons it is possible to divide space in a logical manner, thus increasing positional granularity over that of a single beacon. This process of inference can also be applied to cellular networks. Seeing a particular cell enables a coarse geographic location to be inferred. In terms of granularity the sighting of an 802.11 access point enables a finer positional estimate than that of a cellular base station. However in terms of coverage and user accessibility the cell phone is far more ubiquitous than 802.11; everyone has a cell phone.

Aside from beacon information we are also able to use the signal strength levels of wireless beacons to infer location. This technique known by the term location fingerprinting is possible because, at the same physical location the signal strength levels from a wireless beacon will typically be constant. Therefore matching current signal strength levels with those stored in a radio map enables the distinction between different physical places to be made. This type of information enables the spatial environment to be divided into more zones than using wireless beacon information alone.

We have discussed three sources of information abundant in our daily lives that can be used to infer location. Each source has strengths and weaknesses. Wireless signal strength levels are susceptible to multi-path fades, diffraction and reflection and hence are typically very noisy. Cell phones typically only track up to seven cells at any single time. However in dense urban environments there will potentially be far

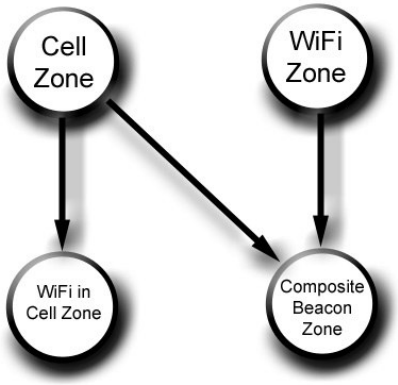



Figure 2: A Bayesian network for fusing cellular, WiFi and beacon positional information.

more than seven cells visible from a single location. Hence the same wireless beacons may not always be observable at the same physical location. However by fusing these sources an increase can be made both in terms of positioning reliability and granularity. In the following section we present a Bayesian network for location inference using this data.

## 4.1 Bayesian Network

Bayesian networks are directed acyclic graphs, where nodes represent random variables and edges or arcs represent the causal relationships between nodes. Each node consists of a set of mutually exclusive states. At each node a probability distribution is defined. Nodes without parents are assigned unconditional probability distributions and those with parents are assigned conditional probability distributions, that is: $P(A_i|B_1,...,B_n)$ where $B_1,...,B_n$ represents the parents of $A$. The joint probability distribution is calculated using the chain rule:

$$P(X_1,X_2,...,X_n) = \prod_{i=1}^{n} P(X_i|Pa_i) \quad (2)$$

where $Pa_i$ represents the parents of $X_i$. By applying evidence at certain variables that is $P(A_i|e)$ where $e$ is evidence, we are able to use the chain rule to determine the probability of an event occurring given limited or partial information. This enables a probabilistic approach to be taken to sensor fusion. For a detailed introduction to Bayesian networks we recommend Heckerman [8].

In Figure 2 we present a Bayesian network that infers the location of a user using the positional data discussed in the previous section. In this network there are two parent nodes and two child nodes. The '*Cell-Zone*' node represents the Cell-Zone that a user is currently located in. This node contains attributes representing all of the different Cell-Zones used in the environment. The number of Cell-Zone attributes is defined by the $k$ value used when clustering the cellular signal strength data. The '*WiFi-Zone*' node is the equivalent of the Cell-Zone node but using WiFi as opposed to cellular signal strength data. As with the Cell-Zone node, the number of WiFi zones is determined by the $k$ value used when clustering the WiFi data. A Composite-Beacon-Zone represents the unique combination of wireless beacons visible at a single point in time. Membership to this type of zone is determined by matching the visible beacons in a measurement sample with those in a given Composite-Beacon-Zone.

We are not agnostic about the type of beacon, cellular of WiFi, we treat the sighting of each of these types individually. Initial experiments found that upon returning to the same position in an environment typically the same 802.11 access points were visible, only those with weak signal strength levels were sighted intermittently. In contrast, the same set of cellular beacons was not always visible at the same physical position. This is due to the limitation of only being able to concurrently monitor seven cells at any point in time.

At any point in time a user will be located in three different zones, a Cell-Zone, a WiFi-Zone and a Composite-Beacon-Zone. This combination of zones forms a qualitative coordinate (Cell-Zone, WiFi-Zone, Composite-Beacon-Zone). By looking at the frequency that a user is placed in a combination of these zones enables inference of the relationships between them. For example, whilst placed in Cell-Zone A a user may notice that their current WiFi zone at this time is typically zone B. Therefore when placed in WiFi-Zone B they can infer they are likely to also be placed in Cell-Zone A. In the Bayesian network this relationship is modelled via the '*WiFi-In-Cell-Zone*' node. The '*Composite-Beacon-Zone*' node represents the equivalent relationship for Cell-Zone, WiFi-Zone and Composite-Beacon-Zone.

In the following section we discuss how the conditional probability distributions can be learnt from historical data and in Section 4.3 we illustrate how to apply evidence to the network to obtain stronger location estimates.

## 4.2 Node Probability Distributions

At each node in a Bayesian network a probability distribution must be defined. For nodes with no parents this distribution is unconditional and for those with parents the distribution is conditioned upon the parent nodes. As with any Bayesian network, probability distributions are populated either by a domain expert or by learning from historical data. We learn these probabilities using the same training data used to create the initial zone based representation of the environment. To populate the Cell-Zone and WiFi-Zone probability distributions we use the Euclidean distance from a measurement sample to a cluster (zone). The closer the measurement sample to the cluster the greater the probability of the user being located in that zone (cluster). We normalise the data by dividing the distance from a given cluster to a measurement sample by the sum of distances from that sample to every other cluster in the environment.

A Composite-Beacon-Zone consists of a unique combination of beacons visible at a single point in time. Membership to this type of zone is determined by matching the visible beacons in a measurement sample with those in the given

Composite-Beacon-Zone. Certain Composite-Beacon-Zones will be visible more often in certain Cell-Zones and certain Wifi-Zones. Hence knowing the current Composite-Beacon-Zone enables a Cell-Zone or WiFi-Zone positional estimate to be made with an increased confidence. This relationship is represented in the Bayesian network by the links from the Cell-Zone and WiFi-Zone nodes to the Composite-Beacon-Zone node. As such the conditional probability table distribution for the Composite-Beacon-Zone is dependent upon the state of the Cell-Zone and WiFi-Zone nodes. The values in the conditional probability distribution are populated by calculating the frequency a given Composite-Beacon-Zone was visible whilst a Cell-Zone and WiFi-Zone combination was also visible.

The WiFi-In-Cell-Zone node, like the Composite-Beacon-Zone represents the relationship of being concurrently located in different types of zone. In this case, the probability of being located in a given WiFi-Zone whilst also being located in a given Cell-Zone. Again this enables increased assertions to be made regarding both the current Cell-Zone and the current WiFi-Zone. The conditional probability distribution is determined by calculating the frequency a given WiFi-Zone was visible whilst a given Cell-Zone was also visible.

In this network the probability distributions for the Composite-Beacon-Zone and WiFi-In-Cell-Zone nodes are static for a given set of training data. The root nodes, Cell-Zone and WiFi-Zone nodes are however dynamic and the unconditional probability distributions are updated with each new positional dependent measurement.

### 4.3 Applying Evidence

In this section we demonstrate how it is possible to make a *stronger* estimate of a users qualitative location by applying evidence to the Bayesian network described in the previous section. We demonstrate how to add evidence to the Bayesian network to determine the following probability:

$$P(CZ = A|WZ = A, WICZ = A, CBZ = A)$$

We start by looking at the probability of being in a particular Cell-Zone given evidence about the current WiFi-Zone, WiFi-In-Cell-Zone and Composite-Beacon-Zone. By using Bayes rule we can write this as:

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WZ, WICZ, CBZ, CZ)}{P(WZ, WICZ, CBZ)}$$

The states of Cell-Zone are mutually exclusive, hence we are able to transform the denominator to give:

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WZ, WICZ, CBZ, CZ)}{\sum_{CZ'} P(WZ, WICZ, CBZ, CZ')}$$

By using the product rule we can now expand both numerator and denominator to give.

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WZ|WICZ, CBZ, CZ) * P(WICZ|CBZ, CZ) * P(CBZ|CZ) * P(CZ)}{\sum_{CZ'} P(WZ|WICZ, CBZ, CZ') * P(WICZ|CBZ, CZ') * P(CBZ|CZ') * P(CZ')}$$

At this point we have an equation that is not representative of the conditional independencies in our Bayesian network. We therefore need to update statements such as $P(WICZ|CBZ, CZ)$ with the relationships shown in Figure 2. This gives:

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WZ) * P(WICZ|CZ) * P(CBZ|CZ, WZ) * P(CZ)}{\sum_{CZ'} P(WZ) * P(WICZ|CZ') * P(CBZ|CZ', WZ) * P(CZ')}$$

We are then able to simplify by removing the common factor $P(WZ)$ from both the numerator and denominator. This is possible because the prior probability for the WiFi-Zone has no direct effect on the Cell-Zone probability. This simplification gives:

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WICZ|CZ) * P(CBZ|CZ, WZ) * P(CZ)}{\sum_{CZ'} P(WICZ|CZ') * P(CBZ|CZ', WZ) * P(CZ')}$$

We are now in a position to determine the value of $P(CZ = A|WZ = A, WICZ = A, CBZ = A)$ by substituting known evidence.

$$P(CZ|WZ, WICZ, CBZ) = \frac{P(WICZ{=}A|CZ{=}A) * P(CBZ{=}A|CZ{=}A, WZ{=}A) * P(CZ{=}A)}{\sum_{CZ=A' \in (yes,no)} P(WICZ{=}A|CZ{=}A') * P(CBZ{=}A|CZ{=}A', WZ{=}A) * P(CZ{=}A')}$$

We can now solve this by substituting the values from the conditional probability tables. This allows us to make stronger estimate of a users position thus increasing positioning system performance. This process is repeated to determine the probabilities for the Composite-Beacon-Zone and WiFi-Zone zones.

## 5 ASSESSING PERFORMANCE

In Section 3.2 we demonstrated how it was possible to construct a zone-based representation of a spatial environment in an unsupervised manner. As part of the process the deployer had to select the number of zones to cover the application environment. As such, a range of values are used with the performance of each *solution* being evaluated and the most appropriate selected. We use the term '*solution*' to refer to both the Cell-Zone and WiFi Zone radio maps together with the Composite-Beacon-Zone map. In this section we discuss the different aspects of performance that need to be considered when selecting a solution.

When considering the performance of a solution we must assess three factors: reliability, granularity and substantiality.

In terms of performance, '*reliability*' refers to consistently positioning a user in the same qualitative zone when they are at the same physical position. The positional granularity of a solution is dependent upon the number of distinguishable or effective zones. We use the term '*effective zone*' to refer to a zone that a user has been identified as being located in. Ideally the number of effective zones will be equal to the total number of created zones. This is not however realistic with all types of positional dependent measurements, particularly noisy sources such as GSM and 802.11 signal strength levels. As such, we remove unused zones and instead only use the effective zones.

Using reliability and granularity metrics alone does not indicate whether a solution is suited for a given environment. For example, a radio map may place an application user in a single zone 90% of the time and, during the other 10%, briefly place a user in each of the remaining zones. In this situation both granularity and reliability metrics may indicate excellent performance. However, the usability of the solution is relatively poor. In a perfect solution, when given a set of training data that has been partitioned into $n$ clusters (zones), *replaying* that data - working out the qualitative location given the measurement sample from the training data - will result in the user being placed in each zone for an equal amount of time ($\frac{1}{n}$). In practice, during clustering some zones 'grow' to contain more data than others and hence are typically *matched* more often. As such, when assessing solution performance, it is preferable to have an indication of how many times a user was placed in a zone for a substantial, identifiable amount of time. We refer to this aspect of performance with the term '*substantiality*'. We assess this aspect using the following function:

$$t = \sum_{j=0}^{N} (\frac{1}{j} - \frac{s}{a}) \tag{3}$$

where $N$ is the number of generated clusters, $j$ is a given cluster, $a$ is the total number of measurement samples in the path and $s$ is the number of times that, during the course of the path, the users qualitative location was $j$. The greater the value of $t$ the poorer the solution has performed in terms of zone substantiality. We use the term *time-error* to refer to the value of $t$. A time-error of zero would indicate that a user was placed in each zone an equal amount of time.

# 6 IN PRACTICE

In this section we discuss the performance of the Bayesian network described in Section 4.1. First, in Section 6.1 we assess performance using controlled errors in simulations. Then in Section 6.2 we discuss performance using real-world data collected in metropolitan environments.

## 6.1 Simulations

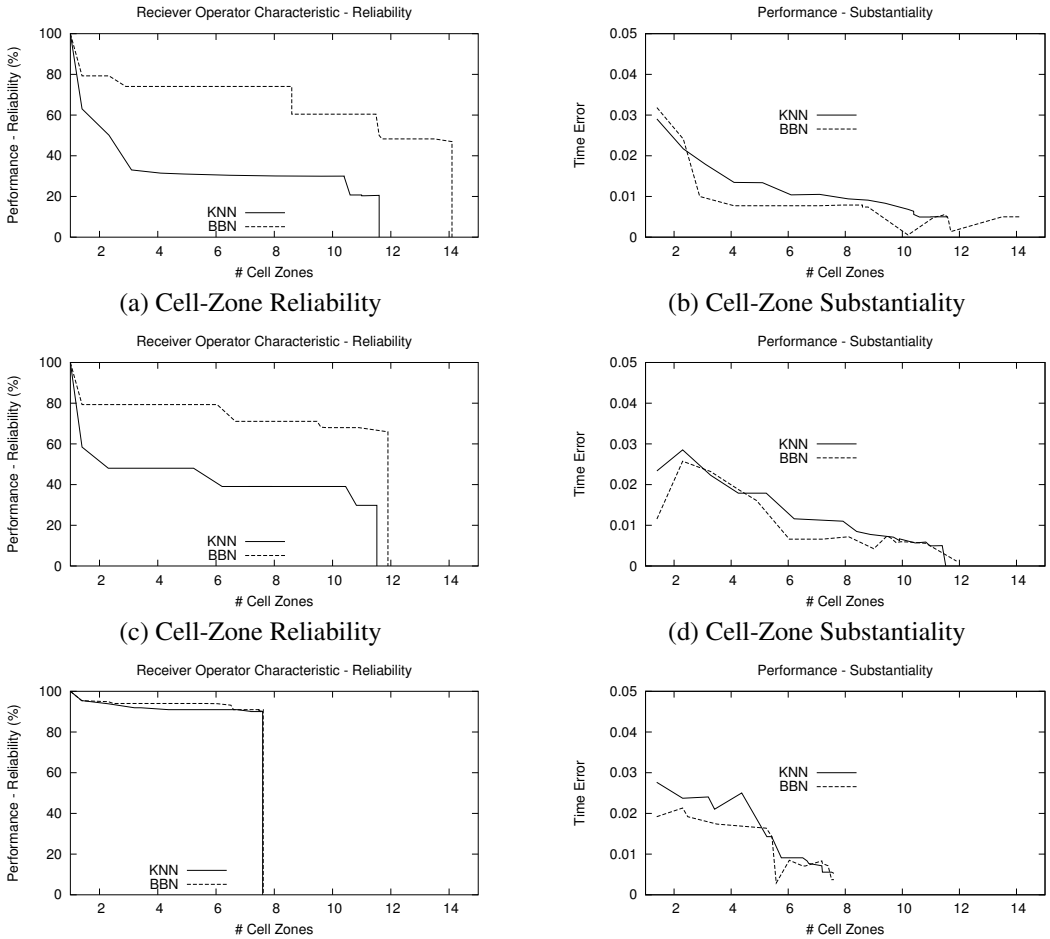
In total we carried out three simulations to assess the effects of noise of each node in the Bayesian network. For each simulation we generated 6000 data-points. Each data-point consisted of seven cell strength readings and a variable number of WiFi signal strength readings. Half of this data was used for training and half was used for testing. The data-points simulated controlled errors. The data-point log files were processed in the same fashion as log files containing real world data.

Figures 3a and 3b show the performance levels of the Bayesian Belief Network (BBN) against the K-Nearest-Neighbour (KNN) when determining a users current Cell-Zone. In simulation one, data was generated to create distinguishable Composite-Beacon-Zones and WiFi zones. Both types of zone shared the same borders. The cellular data was created as noise using random signal strength levels. The aim of this simulation was to see whether it is possible to reliably infer the current Cell-Zone despite noisy cellular data. The results of simulation two are presented in Figures 3c and 3d. In this simulation, clearly distinguishable Composite-Beacon-Zones were created but both WiFi and cellular data were generated as noise using random signal strength values. In the third simulation we generated data-points with clearly distinguishable Cell-Zones, WiFi-Zones and Composite-Beacon-Zones. The performance of KNN and BBN in this simulation are shown in Figures 3e and 3f.

When noise was introduced we found the BBN to offer superior performance in terms of reliability over KNN. When using clear, distinguishable data (at all nodes) we did not find any noticeable performance gain over KNN. The drop in terms of granularity in simulation three is due to the generated data. The data set was structured to create eight distinguishable zones. Hence with this data the maximum number of cell zones was approximately eight. With noisy cellular data but good WiFi and composite beacon data we found that the BBN could extend the granularity of the solution beyond a KNN method. We would have expected KNN to perform at least as well as the BBN because noisy data will typically result in a user being placed in a sporadic fashion across a large number of zones. The BBN should slightly reduce granularity by filtering this noise. With noise at both WiFi and cellular nodes we found that the BBN was still able to produce promising reliability results.

In terms of substantiality in most cases we found the BBN to perform better than KNN. The time-error was typically higher in simulation two than in simulation one for the BBN. This is due to the additional noise introduced at the WiFi node. The time error is high in simulation three despite good data because the data was structured for eight distinguishable zones. Hence, zone numbers lower than this produced in terms of substantiality, poor performance.

These simulations have shown that the Bayesian network can successfully fuse cellular, WiFi and wireless beacon data to infer a qualitative position with an increased confidence. The network can still provide increased performance of KNN even if one or two nodes are providing noisy data. In the following section we discuss performance using real data collected from a metropolitan environment.




(a) Cell-Zone Reliability (b) Cell-Zone Substantiality

(c) Cell-Zone Reliability (d) Cell-Zone Substantiality

(e) Cell-Zone Reliability (f) Cell-Zone Substantiality

Figure 3: Controlled error simulations. Simulation 1, distinguishable Composite-Beacon-Zones and WiFi zones with noisy cellular data. Performance in terms reliability is shown in (a) and substantiality in (b). Simulation 2, distinguishable Composite-Beacon-Zones with noisy WiFi and cellular data. Performance in terms reliability is shown in (c) and substantiality in (d). Simulation 3, distinguishable Composite-Beacon-Zones, WiFi-Zones and Cell-Zones. Performance in terms reliability is shown in (e) and substantiality in (f).

## 6.2 Real World

In order to assess the performance of the Bayesian network using real world data we collected measurement samples from a metropolitan environment. Three volunteers were equipped with Orange SPV C500 cell phones capable of monitoring the signal strength levels for up to seven cells. To obtain 802.11 data users also carried an IPAQ 4700 that ran software to passively scan for WiFi networks. The test area has reasonable GPS coverage, and a GPS receiver was used to collect a ground truth for the samples. Samples were collected once per second. The approach to data collection was deliberately systematic where volunteers were asked to walk along explicit paths. This approach enabled an assessment of the reliability of a users qualitative location to be made. Data was collected at different times of day over an 8-week period in 2005. In total over 85 000 signal strength measurements were taken.

In Figure 4 we show the performance of Bayesian Belief Network (BBN) and K-Nearest-Neighbour (KNN) using real data collected from two different areas of the same metropolitan environment. In the first area WiFi beacons were only visible in approximately 11% of the data-points. In the second area WiFi beacons were visible in approximately 44% of the data-points. We found WiFi data was not as widely available as we had expected. We suspect this is due to the distance from pedestrian paths to nearby buildings. At many points along the path the volunteer was 10-15 metres from the nearest building hence signal strength levels were weak and not always detectable.

In the first area, in terms of reliability, the BBN only performed slightly better than KNN, this is due to the limited WiFi data. In this experiment 89% of the time the BBN was determining location using only cellular data and cellular beacon data. Cellular beacon information is generally very noisy in metropolitan environments with a high number of cell towers provisioning coverage. This is because a typical GSM cell phone can concurrently monitor only 6 neighbouring cells in addition to the current, serving cell. In this environment we found 54 different cells, hence at the same physical position we would *hear* different combinations of cells. Thus reducing the usefulness of cellular beacon information. In terms of granularity, KNN performed better than the BBN. This is as expected. The BBN reduces noise therefore typically places a user in fewer zones than the equivalent KNN approach. In the second area of the environment, the performance gain in terms of reliability of the BBN over KNN was more substantial than in the first area. This is due to the increased availability of WiFi information in the second area. In terms of granularity, both BBN and KNN decline at a faster rate than in area 1. We suspect this is reflective of that area of the environment - fewer distinguishable zones.

In terms of substantiality we found the BBN to typically generate lower time-errors, that is, it placed users in cell zones more equally than KNN. In area one we found that optimum performance was achieved with between four to eight zones. In area two as the numbers of zones was raised above six, the time-error steadily increased. This suggests that this environment should be covered by six or less zones. This is confirmed by looking at the performance reliability in Figure 4a.

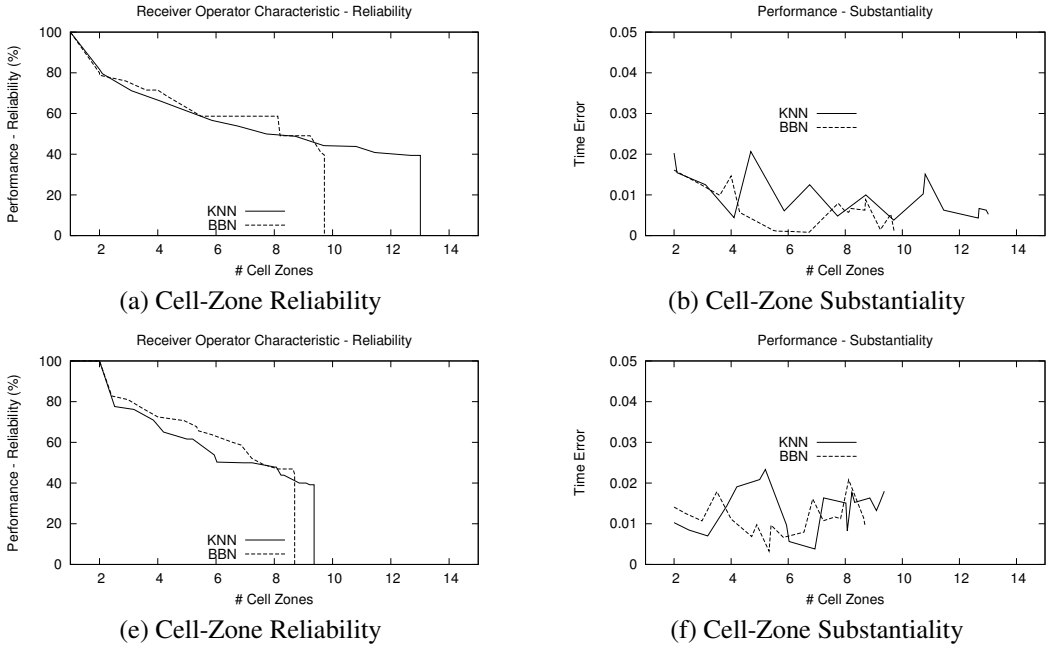
Given these experiments using data collected from the real world, the Bayesian network offers a slight increase in reliability when only cellular and wireless beacon positional information are available. This is due to the noise associated with cellular beacon information, particularly apparent in dense urban environments. Environments where cellular, WiFi and wireless beacon information are readily available benefit most from applying this Bayesian network.

# 7 CONCLUSIONS AND FUTURE WORK

We have designed a qualitative location system that operates by inferring location using positional dependent signals already abundant in our every day lives. We contain measurement and environmental limitations within the location system and return current location in the form of a spatial zone. This approach differs from the more traditional quantitative based positioning systems that return an estimate and variable error. We feel that a qualitative approach is particularly appropriate when using 'noisy' positional data and when aiming to provide wide coverage over environments typically considered 'harsh', such as open outdoor environments and those in built-up metropolitan areas.

In terms of performance, we have found selecting the most appropriate solution for a given environment to be a trade-off between two factors; reliability and granularity. As the number of zones used to represent the environment are increased so to is the ability to support fine-grained location services. However, the reliability of placing a user in the same qualitative zone at the same physical position is reduced. This problem can be alleviated by the introduction of additional sources of position dependent measurements. We have implemented this work using WiFi and GSM signals strength levels and wireless beacon information. In both simulation and real world experiments we found the Bayesian network to offer an increase in performance over a KNN approach.

In this paper we have demonstrated how a qualitative representation of the spatial environment can be constructed in an unsupervised, automated manner after a simple calibration procedure. Given this type of information, deployers of context-aware applications can assess what is feasible given the positional data available to them. For example, consider a context-aware application designed to be used on a shopping street. The application provides the user with details of current product offers when the user is standing in front of a shop. Using the approach presented in this paper the deployer can assess which shop fronts are distinguishable from other shop fronts. They are able to see the limits of the environment and measurement service in terms of the areas of interest, the shop fronts. Currently the calibration phase requires the user to collect positional dependent measurements in the application environment prior to deployment. For the future, we intend to make this a continuous process that is carried out passively at runtime. This will enable a current and useful radio map to be maintained with a minimal overhead.




(a) Cell-Zone Reliability

(b) Cell-Zone Substantiality

(e) Cell-Zone Reliability

(f) Cell-Zone Substantiality

Figure 4: Experiment 1 - limited 802.11 coverage. Performance in terms reliability is shown in (a) and substantiality in (b). Experiment 2 - good 802.11 coverage. Performance in terms reliability is shown in (c) and substantiality in (d).

## REFERENCES


[1] Ian Anderson and Henk Muller. Towards qualitative positioning for pervasive environments. In *the Third International Conference on 'Computer as a tool' (Eurocon 2005)*, November 2005.

[2] Michael Angermann, Patrick Robertson, and Thomas Strang. Issues and requirements for Bayesian approaches in context aware systems. In *LoCA*, pages 235–243, 2005.

[3] Paramvir Bahl and Venkata N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM (2)*, pages 775–784, 2000.

[4] Mauro Brunato and Roberto Battiti. Statistical learning theory for location fingerprinting in wireless LANs. *Computer Networks*, 47(6):825–845, 2005.

[5] Yu-Chung Cheng, Yatin Chawathe, Anthony LaMarca, and John Krumm. Accuracy characterization for metropolitan-scale Wi-Fi localization. In *MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pages 233–245, New York, NY, USA, 2005. ACM Press.

[6] Anthony LaMarca et al. Place lab: Device positioning using radio beacons in the wild. In *Proceedings of PERVASIVE 2005, Third International Conference on Pervasive Computing*, Munich, Germany, 2005.

[7] Andreas Haeberlen, Eliot Flannery, Andrew M. Ladd, Algis Rudys, Dan S. Wallach, and Lydia E. Kavraki. Practical robust localization over large-scale 802.11 wireless networks. In *MobiCom '04: Proceedings of the 10th annual international conference on Mobile computing and networking*, pages 70–84, New York, NY, USA, 2004. ACM Press.

[8] David Heckerman. A tutorial on learning with Bayesian networks. pages 301–354, 1999.

[9] John Krumm and Eric Horvitz. LOCADIO: Inferring motion and location from Wi-Fi signal strengths. In *MobiQuitous*, pages 4–13, 2004.

[10] Andrew M. Ladd, Kostas E. Bekris, Algis Rudys, Guillaume Marceau, Lydia E. Kavraki, and Dan S. Wallach. Robotics-based location sensing using wireless Ethernet. In *Proceedings of the Eighth ACM International Conference on Mobile Computing and Networking (MOBICOM)*, Atlanta, GA, September 2002.

[11] Heikki Laitinen, Jaakko Lahteenmaki, and Tero Nordstrom. Database correlation method for GSM location. In *Proceedings of the 53rd IEEE Vehicular Technology Conference*, Rhodes, Greece, May 2001.

[12] Anthony LaMarca, Jeffrey Hightower, Ian E. Smith, and Sunny Consolvo. Self-mapping in 802.11 location systems. In *the Seventh International Conference on Ubiquitous Computing (UbiComp)*, pages 87–104, 2005.

[13] Julie Letchner, Dieter Fox, and Anthony LaMarca. Large-scale localization from wireless signal strength. In *Proceedings, the Twentieth National Conference on Artificial Intelligence, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 15–20, 2005.

[14] Veljo Otsason, Alex Varshavsky, Anthony La Marca, and Eyal de Lara. Accurate GSM indoor localization. In *the Seventh International Conference on Ubiquitous Computing (UbiComp)*, September 2005.