A Statistical Modeling Approach to Location Estimation

Teemu Tonteri

Helsinki, 25th May 2001

Master's Thesis

University of Helsinki
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Teemu Tonteri

Työn nimi — Arbetets titel — Title

A Statistical Modeling Approach to Location Estimation

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	25th May 2001	53

Tiivistelmä — Referat — Abstract

Location services provide users of cellular telephones with information about their location. In order to implement location services, several location estimation methods have been developed. Some of them, such as the GPS satellite navigation system, require non-standard features, either from the cellular telephone or the cellular network. However, it is possible to use the existing GSM technology for location estimation by taking advantage of the signals transmitted between the cellular telephone and the network. A problem with such solutions is usually their inadequate location estimation accuracy.

The thesis reviews some current location estimation methods. In addition, it describes propagation models, which are used to predict some properties of radio signals, such as signal strength. The use of propagation models in location estimation is discussed. This leads to an approach to location estimation which is different from the prevailing geometric one. We call our approach *the statistical modeling approach*. In the empirical part of the thesis, a location estimation method based on a statistical signal strength model is presented.

Computing Reviews Classification:
G.3 (Probability and Statistics),
J.2 (Physical Sciences and Engineering)

Avainsanat — Nyckelord — Keywords
location estimation, propagation modeling, Expectation–Maximization algorithm

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Teemu Tonteri

Työn nimi — Arbetets titel — Title

Tilastollisen mallinnuksen lähestymistapa paikannukseen

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro Gradu -tutkielma	25. toukokuuta 2001	53

Tiivistelmä — Referat — Abstract

Matkapuhelimien paikannuspalvelut tarjoavat matkapuhelimen käyttäjälle tiedon käyttäjän sijainnista. Paikannuspalvelujen tuottamiseksi on kehitetty useita erilaisia paikannusmenetelmiä. Jotkut niistä, kuten GPS-satelliittipaikannus, vaativat joko matkapuhelimelta tai matkapuhelinverkolta lisäominaisuuksia, joita niissä ei tavallisesti ole. Paikannus voidaan kuitenkin tehdä nykyisellä GSM-teknologialla käyttäen hyväksi pelkästään matkapuhelimen ja tukiasemien välisiä signaaleja. Tällöin ongelmaksi muodostuu paikannustarkkuus, joka on nykymenetelmissä on riittämätön useisiin paikannuspalveluihin.

Tutkielmassa luodaan katsaus nykyisiin paikannusmenetelmiin. Lisäksi kuvataan signaalin ominaisuuksien, kuten voimakkuuden, ennustamiseen käytettäviä malleja ja pohditaan niiden käyttömahdollisuuksia paikannuksessa. Tämä johtaa niin sanottuun *tilastollisen mallinnuksen lähestymistapaan*, joka poikkeaa tavanomaisesta, geometrisesta lähestymistavasta. Tutkielman kokeellisessa osassa esitetään tilastolliseen signaalinvoimakkuusmalliin perustuva paikannusmenetelmä.

Luokitus (Computing Reviews Classification):
G.3 (Probability and Statistics),
J.2 (Physical Sciences and Engineering)

Avainsanat — Nyckelord — Keywords
paikannus, etenemismallit, Expectation–Maximization -algoritmi

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Acknowledgements

I thank everyone who helped in preparing this thesis: first of all, my supervisors Prof. Henry Tirri, and Doc. Petri Myllymäki; secondly, Mr. Tomi Silander who helped in designing the necessary algorithms and in calculus; Mr. Pekka Tonteri for proofreading; co-workers in the Complex Systems Computation Group, and colleagues in the University of Helsinki and the University of Jyväskylä for various discussions; Prof. Steinar Andresen and Mr. Jari Syrjärinne for delivering material; and last but not least, Eira for mental support.

Helsinki, 25th May 2001
Teemu Tonteri

Contents

List of Abbreviations

AMPS Automatic Message Processing System

AOA Angle of Arrival

ASME American Society of Mechanical Engineers

COST European Cooperation in the Field of Scientific and Technical Research

EDGE Enhanced Data Rates for GSM Evolution

E-OTD Enhanced Observed Time Difference

ETSI European Telecommunications Standards Institute

FGCS Federal Geodetic Control Subcommittee

FTSC Federal Telecommunications Standards Committee

GLONASS Global Navigation Satellite System

GPRS Global Packet Radio System

GPS Global Positioning System

GSM Global System for Mobile Telecommunications

MLE maximum likelihood estimate

NMT Nordic Mobile Telephone System

PDA Personal Digital Assistant

p.d.f. probability distribution function

SIM Subscriber Identity Module

SMS Short Message Service

TACS Total Access Communication System

TDOA Time Difference of Arrival

TOA Time of Arrival

UMTS Universal Mobile Telecommunication System

WAP Wireless Application Protocol

Chapter 1

Introduction

The popularity of cellular telephones has increased tremendously during the recent years, and the trend doesn't seem to be slowing down. In August 2000 the number of cellular telephone users in the world was about 570 millions. With the average rate of a quarter of a million new subscriptions *per day*, the total number of subscribers is expected to reach one billion in only a couple of years (UMTS Forum, 2000). These numbers are remarkable for such a young technology: the first cellular telephones were introduced not earlier than in the late 1970s. After two decades, at the turn of the millennium, cellular telephones have developed enough to compete with fixed-line telephones. Recent advances in digital cellular technology allow transmission of not only speech, but also text and data with speed comparable to fixed-line networks. In this introductory chapter we shall take a cursory look into the basics and history of cellular technology, and discuss some of its novel applications.

A fundamental concept of cellular telephone networks is a *cell*.[1] The operating area of a cellular network is divided in cells, each of which is associated with one base station. For simplicity, the cells are often presented as if they had a hexagonal form, as shown in Figure 1.1, although in real life, they overlap each other in a way that allows mobile units to communicate with several base stations in most locations. The idea of cellular layout is to allow efficient use of bandwidth: in GSM systems, for instance, each cell is allocated a group of frequency bands which is completely different from the group allocated to the neighboring cells. This way two cells can operate on the same frequency without interfering with each other, as long as the cells do not overlap.

Without doubt, mobility is the most important feature of cellular telephone systems. Small-scale mobility is provided by the cordless radio inter-

[1] See (FTSC, 1996) and (Rappaport, 1996) for definitions of telecommunications terms.

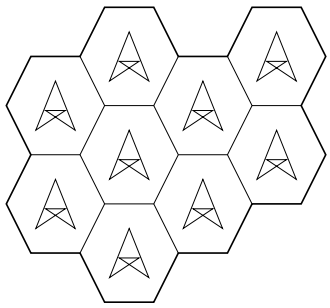

Figure 1.1: An idealized diagram of a cellular network layout.

face between base stations and mobile units. Allowing mobile units to access the network through any base station in the network results in large-scale mobility. Moreover, modern cellular systems perform *hand-overs*, i.e. change the active base station on the fly without interrupting an active call.

During the 1980s several countries adopted analog cellular networks, in the United States AMPS, and in some European countries and in the Nordic countries TACS and NMT respectively (Rappaport, 1996). At that time the systems were often disabled due to lack of sufficient network coverage and capacity. Transmission of other than speech was not considered necessary, and it would have been difficult with the low-performance analog technology. The first digital GSM cellular network started its operation in 1991. Currently GSM network have been installed in more than 100 countries, and the shift from analog to digital technology is accelerating. Digital technology provides an easy way to transmit textual information, and this possibility is exploited by the SMS technology allowing users of GSM handsets to exchange text messages. The transmission speed of a GSM network is sufficient for speech communication and exchange of SMS messages, but the need for data transmission drives the development towards technologies with higher transmission speed.

The next step from GSM will be GPRS. The first commercial GPRS network started operating in the end of the year 2000. Current GSM networks can be upgraded to GPRS with additional network components. Thus its cost compared to building a completely new network infrastructure is small. The effective data transmission speed of GPRS is expected to be between 20 and 65 kilobits per second, which is an improvement over 9.6 kilobits of GSM networks. What is perhaps even more important than increased transmission speed, is that unlike in GSM, a packet-switched protocol is used in GPRS. In practical terms, GPRS is better suited for transmission of text, binary and graphical data than GSM, and it allows billing based on the amount of

transferred data. The latter feature is essential to interactive applications, because no cost should assigned to the time the user views the response. In contrast, in GSM systems billing is always based on connection time.

Until the recent years the development of cellular telephone handsets has primarily focused on decreasing the size of the devices and increasing battery capacity. Lately the emphasis has been shifting to improving the functionality of the devices. Some recent products anticipate the symbiosis of cellular telephones and PDAs with their e-mail, notebook and calendar features. The SIM Application Toolkit (ETSI, 1998) and WAP (WAP Forum, 1998) standards mark milestones in the evolution towards multi-functional telephones. Both of them allow operators and third parties to develop specialized applications for GSM handsets, without the necessity to modify the handset.

All these technological advances would be of little use without co-evolution of the ways in which they are exploited. Among the most interesting ways to exploit the possibilities of modern communication technology, is the concept of *location-aware* computing: devices which can be located or which can locate themselves, and services based upon them. These so called location-based services have great potential in areas such as personal security, navigation, tourism, and entertainment. The most obvious location-based service is one answering questions like "Where am I?", and "Where is the nearest shop/bus-stop/hospital?". Now that graphical and interactive applications are technically feasible, it would even be easy to implement an application which presents a map labeled with a mark saying "You are here". Secondly, location can be thought as a filter for the ever-increasing amount of information available to us every day. People usually don't want to know about daily offerings of supermarkets, let alone of those which are located hundreds of kilometers away.

Location information can be useful for other people than the user of the location-aware device as well. For instance, people often want to know where their friends are, companies want to know where their delivery vehicles are, rescue officials want to know where injured people are, and so forth. In the United States location-based services, and in particular location of the origin of emergency calls has been considered so important, that it is becoming obligatory for the local network operators to provide means for it. This so called Enhanced-911 requirement is scheduled to become effective in October 2001. Similar actions have been considered in the European Union as well.

At this point one may have a concern about possible illegal and unethical use of information concerning individuals' whereabouts. It seems that scientists and engineers are not expected to bother their minds with considerations of the goods and evils of the technological advances they pursue; it

was once said by a certain physicist[2] that

> " When you see something that is technically sweet, you go ahead and do it and you argue about what to do about it only after you have had your technical success. "

However, regulative actions are being carried out in order to prevent such problems. With this in mind, we turn our look back to technical considerations.

The location of a cellular telephone can be estimated using radio signals transmitted or received by the telephone. Some location estimation methods, such as GPS, are based on signals transmitted from satellites, while others rely on terrestrial communication. Additional costs to the service provider are minimal in systems based on existing cellular network infrastructure. However, the location estimation accuracy of such systems is often inadequate for many location services. Improving the accuracy of location estimation systems based on the existing cellular network infrastructure would be very useful. It is the main motivation of this thesis.

One of the most severe problems facing cellular telephone systems is the complex propagation of radio waves in environment with obstructions and reflecting objects. In order to ensure good coverage in their cellular networks, operators use so called cell planning tools which are based on radio wave propagation models. Such models use information about the environment and combine it with knowledge about phenomena such as signal attenuation, reflection, diffraction and interference. The dependency between the location of the receiver and observable signal properties is important for location estimation as well. Despite this fact, the fusion of propagation models and location estimation is rarely mentioned in the literature.

Radio wave propagation, and location estimation are dealt in great detail in the literature. These topics are the concern of the first chapters of the thesis: Chapter 2 presents the principles of radio wave propagation and propagation models, and in Chapter 3 we describe some existing location estimation systems. After reviewing relevant literature, we will discuss location estimation from a point of view which is different from the traditional, geometric one. In particular, a location estimation system based on a statistical propagation model will be proposed in Chapter 4. Finally, some concluding remarks are presented in Chapter 5.

[2](USAEC, 1954) as quoted in (ASME, 2000)

Chapter 2

Radio Wave Propagation

With location estimation systems based on radio signals, it is important to know the propagation properties of electromagnetic radiation. Phenomena, such as signal attenuation, reflection, scattering and diffraction have important roles in location estimation. Their importance is emphasized in non-satellite systems which have to operate in complex propagation environments, such as urban or mountainous areas. This chapter addresses the most important theoretical aspects of radio wave propagation and reviews some propagation models based on them.

2.1 Principles

The basic concept in the theory of electromagnetic radiation is an *electric field*, which is always related to electric current (see Asimov, 1966). An electric field E is defined by its direction and magnitude at each point. The magnitude, denoted by $|E|$, is measured in units of volts per meter (V/m). Periodic fluctuations of an electric field are called *radio waves*. Radio waves can be decomposed in orthogonal components, typically the horizontal and the vertical component. The ratio of the magnitudes of the two components—or equally: the direction of the electric field—defines the *polarization* of the wave (see NAWC, 1997). For instance, if the magnitude of the vertical component is always zero, i.e. the direction vector is always parallel to the horizontal axis, the wave is said to be horizontally polarized.

An electric field corresponds to a *power density flow* F, measured in watts per square meter (W/m^2), which is proportional to the square of the magnitude of the electric field. Given the power density flow, the gain of a receiving antenna, G_r, which depends on the physical size of the antenna and frequency, the wave length λ, and the system hardware loss, L, the received

power is given by

$$P_R = \frac{FG_r\lambda^2}{4\pi L}. \tag{2.1}$$

Even though the wave length λ appears in Equation (2.1), it does not follow that the received power would increase proportionally to the square of the wave length, because the wave length also affects the gain of the receiving antenna G_r. In fact, if the physical size of the antenna and the power density flow are constant, the wave length terms cancel each other out, and thus the received power is independent of the frequency. However, the frequency can indeed affect the power density flow due to interactions with the propagation medium. This issue will be discussed in the following sections.

Because the values of received power vary over a wide range, it is convenient to use logarithmic scale. A ratio of two quantities can can be presented in decibels (dB) which indicates the logarithm of the ratio multiplied by ten. The unit of decibelwatt (dBW) is the ratio of power referenced to one watt. Conversions between watts and decibelwatts are made with the following two equations[1]:

$$P\,[\text{dBW}] = 10\log(P\,[\text{W}]), \tag{2.2}$$

$$P\,[\text{W}] = 10^{\frac{P\,[\text{dBW}]}{10}}. \tag{2.3}$$

For instance, 0 dBW is equal to one watt, 10 dBW is equal to 10 watts, 20 dBW is equal to 100 watts, etc. The unit of decibelmilliwatt (dBm) is defined similarly as the ratio of power referenced to one milliwatt. Conversions between two decibel units, for instance, decibelwatts and decibelmilliwatts, can always be performed simply by adding a constant to the original value. The following two equations are used for converting decibelwatts to decibelmilliwatts and *vice versa*:

$$P\,[\text{dBm}] = P\,[\text{dBW}] + 30, \tag{2.4}$$

$$P\,[\text{dBW}] = P\,[\text{dBm}] - 30. \tag{2.5}$$

Because of the simple relationship between different decibel units, we will hereafter simply use the word decibel to refer to any kind of decibel unit. In such cases all decibel quantities must be expressed in the same units.

[1] The notation "$\log(\cdot)$" is used to denote base 10 logarithm, while the natural logarithm is denoted by "$\ln(\cdot)$".

2.1.1 Free-space Attenuation

Because a wave front proceeds in three dimensions, the maximum received power at distance d must decrease in the inverse of the area of a sphere with radius d. If the absorption loss of the propagation medium is ignored, the power density flow, F, is given by

$$F = \frac{P_T G_T}{4\pi d^2}, \tag{2.6}$$

where P_T is the transmitted power, G_T is a factor depending on the transmitting antenna, and d is the distance (see Rappaport, 1996; Walke, 1999). Combining Equations (2.1) and (2.6) gives the received power, which is usually given in decibels:

$$\begin{aligned} P_R\,[\mathrm{dB}] = {} & P_T\,[\mathrm{dB}] + 10\log(G_T) + 10\log(G_R) + 20\log(\lambda) \\ & - 20\log(d) - 22.0. \end{aligned} \tag{2.7}$$

Equations (2.6) and (2.7) are valid only in free-space environment, where there are neither reflections, absorption, diffraction nor other distortions. If the line-of-sight between the transmitter and the receiver is obstructed, the received signal power is significantly lower than the free-space equations suggest. Furthermore, they do not necessarily give a good approximation even in line-of-sight conditions.

2.1.2 Absorption

In any real-world communication system, the signals propagate in some medium. In wireless terrestrial systems the medium is mainly the atmosphere and, in lesser degree, materials such as glass, concrete, wood, etc. Due to interactions with the medium, the signal loses a certain proportion of its remaining energy on every unit of distance it propagates. Thus, absorption causes the power density flow to decrease proportionally to γ^{-d}, where d is the distance, and γ is a constant depending on the properties of the medium and signal frequency. This means that in decibel scale, the loss is linear with respect to the distance.

Absorption loss is particularly great in the upper microwave region, where the frequencies are above 10 GHz. With these frequencies the absorption due to atmosphere becomes comparable to the free-space attenuation, especially in heavy rain conditions and with long transmitter–receiver distances (Xu *et al.* , 2000). With frequencies used in most wireless communication systems, below 10 GHz, the atmospheric absorption is insignificant with distances up to 10 km.

Absorption caused by other media than air is generally very strong. Moreover, in addition to absorption, obstructions cause the wave to be reflected, which further decreases the amount of energy passing through. Taking into account both reflection and absorption, the total attenuation per obstruction is typically 1–20 dB below 10 GHz, and 1–60 dB above 10 GHz (Rappaport, 1996).

2.1.3 Reflection

Reflection occurs when a wave meets an obstacle with size much bigger than the wave length. The part of the wave that is not reflected back loses some of its energy by absorbing to the material and the remaining part passes through the reflecting object. In terrestrial communication systems the waves usually reflect from ground, producing a two-ray path between the transmitter and the receiver, shown in Figure 2.1. The *plane of incidence* is defined as the plane containing both the incident ray and the reflected ray, and the *angle of incidence* is the angle between the reflecting surface and the incident ray.

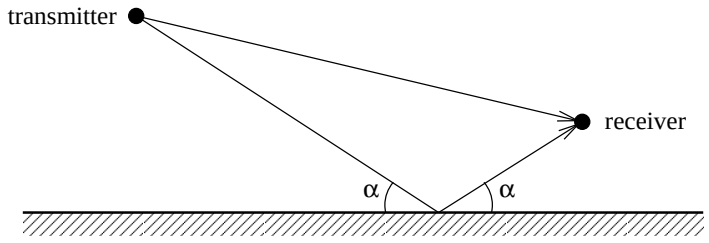

Figure 2.1: Two-ray ground reflection model.

The received signal consists of the direct line-of-sight ray and the reflected ray. The two rays arriving to a receiver can have different phase and in the worst case they cancel each other out. The magnitude of the reflected signal depends on the *Fresnel reflection coefficient*, which depends on the properties of the reflecting ground, the frequency of the wave, and the angle of incidence. Roughness of the reflecting surface causes the propagating waves to *scatter* in all directions, and therefore, the reflection coefficient of a rough surface is smaller than the one of an otherwise identical but flat surface. In general, the reflection coefficient is different for the vertical and the horizontal component of the wave. In such cases, reflection can change wave polarization.

Figure 2.2 presents the attenuation curve of the two-ray model with certain parameters. The exact equation corresponding to the two-ray model is given in (Rappaport, 1996, p. 87). It can be seen from the figure that with

long distances the two-ray model coincides with the *fourth-power approximation*, which is given by

$$P_R\,[\mathrm{dB}] = P_T\,[\mathrm{dB}] + 10\log(G_T) + 10\log(G_R) - 40\log(d) - 22.0, \tag{2.8}$$

where the received power is proportional to the inverse of the *fourth* power of the distance rather than the square of the distance which appears in the free-space model.

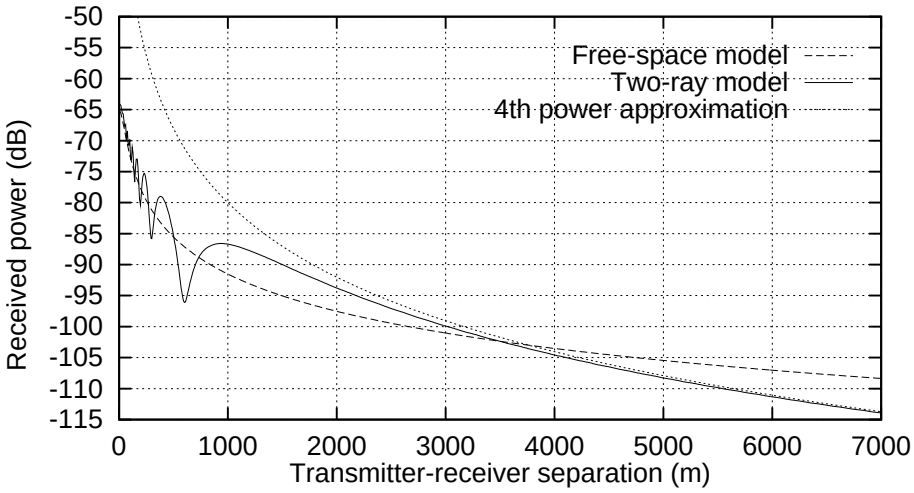

Figure 2.2: The received power referenced to the transmitted power as a function of the transmitter–receiver distance according to the free-space model (Equation (2.7)), the two-ray model (Rappaport, 1996, p. 87), and the fourth-power approximation of the two-ray model (Equation (2.8)). The parameters are: transmitter elevation = 50 m, receiver elevation = 2 m, frequency = 900 MHz, relative permittivity of the ground = 15, antenna gains and system loss = 1.0 (no loss).

2.1.4 Diffraction

According to Huygen's principle, all points on a wavefront are point sources of secondary waves propagating to all directions. Therefore, each time a radio wave passes an edge such as a corner of a building the wave "bends" around the edge and continues to propagate into the area shadowed by the edge. This effect is called *diffraction*. In Figure 2.3 the transmitter is situated near an obstacle. The arrows describing the direction of propagation indicate how

the signal reaches the areas around the corner due to a source of secondary waves situated at the corner of the obstacle. Note that the single source of secondary waves shown in Figure 2.3 is only one of the infinite number of such sources on the wavefront.

The more the waves have to bend around a corner, the more they lose their energy. Therefore the areas to which the rays have to bend more, gain relatively less additional field strength than the areas to which the rays can proceed almost linearly. The field strength of the secondary waves is much smaller than the one of the primary waves. In practice the diffracted waves can be neglected if there is a line-of-sight between the transmitter and the receiver.

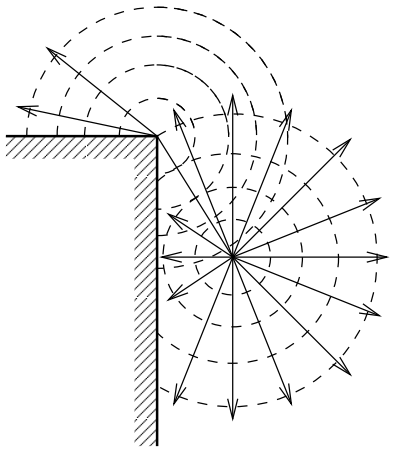

Figure 2.3: Diffraction.

2.2 Propagation Models

Prediction of radio wave propagation is useful in activities such as allocation of bandwidth, cell planning and location estimation. Propagation models are used to predict the properties of the propagating waves, usually the received signal power and its variability. It is also possible to predict polarization, time dispersion, frequency selectivity and other properties that affect the performance of communication systems (Damosso & Correia, 1998; Fleury & Leuthold, 1996).

The theoretical aspects mentioned above can be taken into account on various levels of abstraction, depending on the amount of available information about the environment, and the required accuracy of the predictions.

For instance, when planning satellite communication or radio links spanning tens of kilometers, sufficient accuracy is often reached by taking into account free-space attenuation, absorption and ground reflection. On the other hand, in urban areas reflections and diffraction caused by buildings and scattering caused by trees have a strong influence on wave propagation. Based on how much information about the environment the models use, they can be divided into the categories of general and site-specific models. The division can be further refined as shown in Figure 2.4. For descriptions of the individual models listed in the figure, see (Damosso & Correia, 1998; Andersen *et al.* , 1995; Rappaport, 1996; Wölfle & Landstorfer, 1999). The different groups of propagation models are discussed in the following.

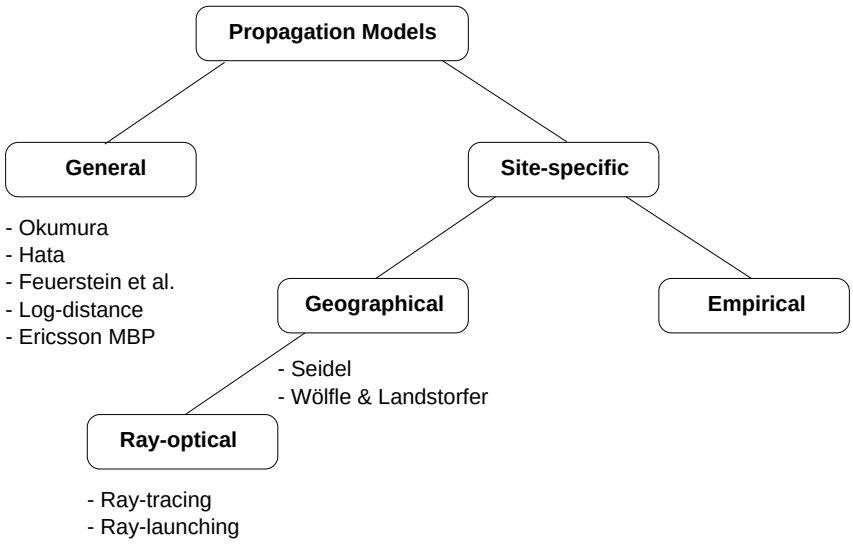

Figure 2.4: Classification of propagation models.

2.2.1 General Models

General models typically describe the field strength as a function of the distance between the transmitter and the receiver. The two-ray ground reflection model and its approximations described in the previous section are examples of general models. The models usually include some factors such as ground properties, or the Fresnel reflection coefficient in the two-ray model. One of the most popular general models is the Okumura model (Rappaport, 1996). It consists of the free-space loss (Equation (2.7)) and a correction factor. The correction factor is given by a function of the distance and

the frequency of the signal. Okumura presented the functions graphically as curves. Different curves exist for open, quasi-open, suburban and urban area. Later, Hata used the empirical data presented by Okumura and gave mathematical formulae known as the Hata model, which closely fits Okumura's curves. The Hata model was extended by the COST-231 working committee to higher frequencies in order to include the 1800 MHz frequency band used in some cellular telephone systems. A drawback of the Hata-Okumura models is that they are restricted to distances of one kilometer or more.

The log-loss (a.k.a. log-distance) model is an extension of the basic free-space attenuation of Equation (2.7). The received power is given by

$$P_R\,[\mathrm{dB}] = P_T\,[\mathrm{dB}] + \beta_0 + \beta_1 \log(d) + e, \tag{2.9}$$

where β_0 and β_1 are parameters indicating how the signal strength decreases as a function of the distance, and e is an error term. If the value of β_1 is -20 the attenuation corresponds to the free-space model, and with the value -40 it corresponds to the fourth-power approximation.

2.2.2 Geographical Models

The propagation of radio waves is well-known on macroscopic level. If geographic information such as earth topography, land use maps, building data, etc., is available, one can use *geographical models* to predict propagation. Such models can be used before the communication system is implemented, which is necessary whenever the predictions are used for optimization of the performance of a system yet to be build. For instance, the so called *ray-optical models* use reflection and diffraction equations to model the paths of the signal, see e.g. (Wölfle & Landstorfer, 1999; Athanasiadou *et al.* , 2000; Corazza *et al.* , 1996). There are two main approaches to ray-optical models: *ray-launching* and *ray-tracing*, see Figure 2.5. In the former several rays are "launched" from the transmitter to all directions. The rays proceed straight on until they hit an obstacle creating one or more reflected or diffracted rays. A prediction of the resulting field is computed by considering at each point all the rays that have passed through the point. Alternatively, in the ray-tracing approach, one starts from some point in the prediction area, and considers potential rays arriving to that point from all directions.

When no empirical data is available, the ray-optical approaches produce the most accurate predictions. However, their use if often prohibited by the lack of detailed and up-to-date information about the environment, and the fact that they are very time-consuming. Some preprocessing techniques are proposed to manage the latter problem (Hoppe *et al.* , 1999). In addition to ray-optical models there are also other geographical models. For

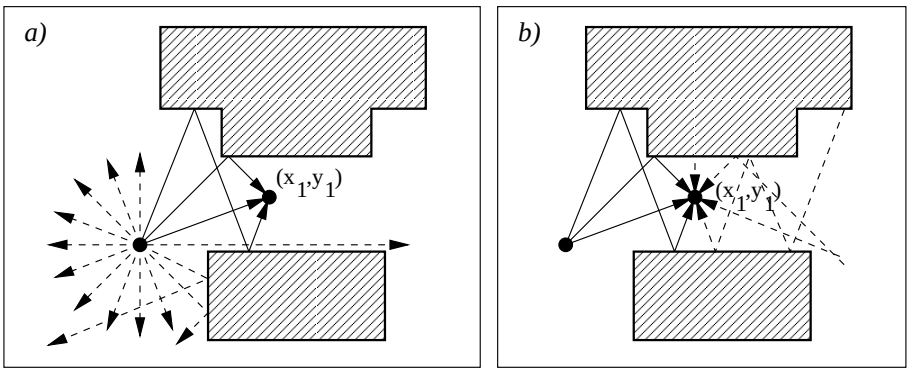

Figure 2.5: Ray-optical models: (a) ray-launching, and (b) ray-tracing. The rays which affect the field at point (x_1, y_1) are denoted by solid arrows. The transmitter is denoted by the black dot on the left side of the obstruction.

instance, neural networks have been used to predict the signal strength inside buildings using some attributes derived from the floor-plan as inputs to the network (Wölfle & Landstorfer, 1998, 1999).

2.2.3 Empirical Models

Empirical models use data collected in the same location where the model is used. Therefore, in contrast to general and geographical propagation models, empirical models can be used only when the system is already in operation. The data is used in a way similar to statistical inference—in fact, many empirical models are statistical— to obtain information on the parameters of the signal in different parts of the area. Empirical models are potentially very accurate, because their predictions[2] correspond to the actual propagation phenomena even when no information about the environment is available. Disadvantages of empirical models are that collecting the measurements can require a lot of work, that they can not be exported to other areas than the one from which data was collected, and that they need to be reconstructed every time the environment changes.

[2]In this context the word *predict* is used in the statistical sense, referring not necessarily to prediction of the future but to prediction of any information that is not known by the predictor.

Chapter 3

Location Estimation Methods

Obtaining the location of a mobile unit is called *location estimation*. Synonymous terms include: *radio location, radio navigation, position location, positioning*, and so forth. Several location estimation systems are reviewed in (Rappaport *et al.* , 1996) and (Syrjärinne, 2001). A vast majority of applications of location estimation use the GPS satellite navigation system which provides location estimates with an accuracy of a couple of meters. Alternatives to satellite-based systems are developed to avoid problems, such as lack of coverage between high buildings and indoors, and high energy-consumption of the devices. These techniques use signals between the mobile unit and terrestrial transmitters or receivers. The transmitters (or receivers) can be either dedicated for this purpose or be a part of a communication system, such as a cellular telephone network.

By using the location of the serving base station of a cellular network, one can easily obtain a very rough location estimate. More accurate estimates require measuring the strength, time delay, angle of arrival or other properties of the signals transmitted between the mobile unit and the base stations. In the geometric approach to location estimation the measurements are transformed into distance and angle estimates. Non-geometric approaches are possible but unusual. However, we argue that they have certain advantages over the geometric ones, and we will return to this question in Chapter 4. This chapter presents the principles of location estimation with focus on the geometric approaches.

3.1 Architectures

There are several ways to distribute the location estimation process between the mobile unit and the other components of the system. First of all, the

observed signals can be sent to the mobile unit (downlink) or by the mobile unit (uplink). In addition, the component performing the actual location estimation can be different from the one that observes the signals. Different combinations of solutions to the above design choices correspond to four main architecture types, discussed below.

In *mobile-based* architectures the mobile unit performs the necessary measurements of downlink signals to infer its own location without any uplink communication. In order for this to be possible the network has to broadcast some assistance data, such as the locations of the base stations.[1] If the assistance data is provided as a point-to-point transmission requested by the mobile unit, the architecture is called *network-assisted*. Unlike in network-assisted architectures, capacity is no problem in mobile-based architectures.

If the mobile unit performs the measurements and transmits the results to the network to be processed, the architecture is *mobile-assisted*. Architectures in which the network receives signals from the mobile unit and performs the necessary operations to estimate the mobile unit's location are called *network-based* architectures. Perhaps the most severe problem of network-based architectures is the so called *hearability* problem: the mobile unit adjusts its transmission power in order to ensure that the active base stations receives its signal with minimal energy consumption. Therefore the other base stations do not necessarily receive the signal as required.

3.2 Algorithms

Location estimation can be performed using several different kinds of measurements. Traditionally the approaches have been geometric: the measurements are transformed into distances or angles with respect to a group of reference points, for instance base stations. A location estimate is then derived using basic geometry. The number of independent measurements related to different reference points depends on the used algorithm. Geometric approaches are not directly able to use more than the minimum number of measurements, although this would in most cases improve the accuracy. Therefore many variations, most of them *ad hoc*, are proposed. The basic geometric algorithms are presented in the following sections. Some examples of non-geometric approaches are presented in (Latapy, 1996; Willassen, 1998; U.S. Wireless, 2001).

[1] Such broadcasts would not be a technical problem in cellular networks. However, the operators are often very secretive about the layout of their base stations.

3.2.1 Angle of Arrival

In the Angle of Arrival (AOA) location estimation method the direction of the signal arriving from the mobile unit to two base stations is measured (Rappaport *et al.* , 1996). If the location of the base stations is known, one can use triangulation to infer the location of the mobile unit. If more than three angle measurements are available, they are not necessarily compatible due to angle measuring errors, and one has to apply more complicated means to obtain a location estimate. Angle measurements require additional hardware, such as antenna arrays, to be installed to the network.

3.2.2 Timing-based Algorithms

If the time delay between transmitting and receiving a signal is known, one can estimate the distance by multiplying by the speed of light. Three distance estimates can be used to estimate location with the Time of Arrival (TOA) method. It is obvious that even a small error in the clock at either the transmitting or the receiving end causes a major error in the distance estimate.

Usually the mobile unit can not be synchronized accurately enough to directly obtain the time used by the signal to travel between a mobile unit and a base stations. One solution to get around the synchronization problem is to use differences in the time delays of several base stations instead of absolute times. Time differences are used in the Time Difference of Arrival (TDOA) method. The basic TDOA equation is

$$r_{i,j} = \sqrt{\left(x_i - x\right)^2 + \left(y_i - y\right)^2} - \sqrt{\left(x_j - x\right)^2 + \left(y_j - y\right)^2}, \tag{3.1}$$

where $r_{i,j}$ is the difference in the time delays between the mobile unit and base stations i and j, (x_i, y_i) and (x_j, y_j) are the coordinates of base stations i and j, and x and y are the coordinates of the mobile unit. Equation (3.1) defines hyperbolic curves and three base stations are required for a location estimate.

If measurements corresponding to more than three base stations are available in either TOA or TDOA methods, the incompatibility problem mentioned in conjunction with the AOA method, can arise. In such cases special heuristics have to be applied to obtain a location estimate. The TDOA method is successfully applied in the GPS system and a version of it for GSM networks is standardized with the title Enhanced Observed Time Difference (E-OTD) (Rantalainen & Pickford, 1999).

3.2.3 Signal Strength

If the signal strength is known, the distance can be estimated in a way similar to one used in the TOA method. Therefore, TOA algorithms are applicable to signal strength measurements. Some non-geometric algorithms based on signal strength measurements have been presented as well (Latapy, 1996; Willassen, 1998). It has been suggested that signal strength is not sufficient for accurate location estimation, meaning that accuracies below a few hundred meters are not achievable (Syrjärinne, 2001).

3.3 Existing Location Systems

Some existing location systems, based on either satellites or cellular networks, are reviewed in the following. Most of the systems are commercial products and information about them is mainly provided by companies which license them. Therefore, the credibility of the information presented below is questionable.

3.3.1 Satellite-based Location Systems

Satellite-based location systems, often called *satellite navigation systems*, and among them especially GPS, are used extensively in military and commercial applications, such as vehicle tracking, navigation, and clock synchronization. Advantages of the GPS system include worldwide availability, high accuracy and the fact that it is free for everyone. GPS is applicable only in areas where there is a simultaneous line-of-sight to several satellites. This rules out the use of conventional GPS systems, for instance, under dense foliage, between high buildings and indoors. In addition, it requires relatively expensive hardware in handsets.

The Russian satellite navigation system, GLONASS, is in principle similar to GPS. However, its functionality is currently limited because there are only nine of 24 satellites operating (Bretz, 2000). The European Union is also planning a satellite navigation system, called Galileo. According to the current plan, Galileo will start operating between 2005 and 2008.

Until recently the accuracy of plain GPS location was about 100 meters. More sophisticated equipment used prelocated reference points in the so called *differential GPS* scheme to obtain accuracy of approximately one meter. However, in the beginning of May 2000 the United Stated Department of Defense removed the so called *selective availability* from the satellite signals, thus eliminating the main source of inaccuracy of GPS for civilian

users. Since then the accuracy of plain GPS without differential corrections has been a couple of meters (FGCS, 2000).

Some GPS systems, such as SnapTrack (SnapTrack, 2001) and Tidget (Belle *et al.* , 1997), use only a sensor recording a "snapshot" of the GPS data and transmit it to a server to obtain a location estimate. SnapTrack uses a network of supporting GPS receivers to demodulate the satellite navigation message, thus being able to use the supporting receivers to perform the task which requires the highest signal level in the GPS location estimation procedure. The supporting receivers also provide aiding data to the mobile unit, enabling it to extract the necessary information from a weaker signal than conventional GPS receivers. SnapTrack has reported accuracy of 3–100 meters even inside buildings and severe blockage and multipath conditions, which would be a remarkable improvement compared to conventional GPS systems.

3.3.2 Network-based Location Systems

Many location estimation systems that are based on the signals transmitted between a cellular telephone and the network are proposed. These so called *network-based location systems*[2] are used in order to avoid the necessity to integrate GPS hardware to handsets or to serve as fall-back systems in locations where GPS is not available. Many network location systems use special receivers to monitor signals from cellular phones. The receivers can be placed either at base stations or separate sensor stations. Systems based on time measurements are usually synchronized with GPS receivers or high-precision clocks.

Implemented location estimation systems using TDOA implemented with additional hardware include Cellocate (Cell-Loc, 2001), TruePosition (TruePosition, 2000), and Cursor (CPS, 2001). Nokia has also implemented a test version of its TDOA system (Ruutu, 2000). Additional hardware is used also by GeoPhone (Radix Technologies, 2001) which is based on a TDOA/AOA hybrid, Telesentinel (KSI, 2001) using AOA, and RadioCamera (U.S. Wireless, 2001; Driscoll, 1998). RadioCamera is an empirical system which monitors the mobile unit's transmission to obtain "fingerprints" of the multipath characteristics of the signal. The fingerprints are compared to a database of fingerprints with known coordinates. The system has been reported to achieve accuracy of 86 meters in 67 percent of location requests in both urban and rural environments as well as in outdoor and indoor conditions (U.S.

[2]Note that the term *network-based location system* does not imply a network-based *architecture*.

Wireless, 1999).

Current systems requiring no additional hardware either to the mobile unit nor the network components have severe accuracy problems. By using the serving base station's location as a location estimate yields an average error of about one kilometer or more, depending on the density of the base stations. A location estimation system developed by ModelSoft uses signal strength measurements made by the mobile unit (ModelSoft, 2001). In both rural and urban areas 95 percent of the location estimates are said to be within 1000 meters of the true location. The system will be implemented in GSM handsets manufactured by Benefon and it will be supported by a Finnish GSM carrier Radiolinja (ModelSoft, 1999). CellPoint (CellPoint, 2001) and Alcatel (Kelsey Group, 2000) have also developed location estimation systems requiring no additional hardware. The required software modifications are implemented with the SIM Application Toolkit technology.

Chapter 4

Statistical Location Estimation

In this chapter a statistical approach to location estimation is presented. The basic idea is to construct a statistical model which describes the distribution of signal strength at any given location, and to use the model to estimate the mobile unit's location when the signal strength is observed. In fact, the model in question is a sort of a propagation model and therefore propagation modeling is strongly linked to this approach. The use of a statistical model allows certain theoretically and practically feasible solutions in the location estimation phase. The approach is different from most of the other location estimation systems presented in the literature.

The chapter is organized as follows: We shall first give a detailed description of a suitable propagation model and show how its parameters can be estimated from empirical data. The rest of the chapter deals with location estimation using the model. The elementary probability theory and statistics used in this chapter can be found in, for instance, (DeGroot, 1986).

4.1 Propagation Model Description

A propagation model predicts some properties of a radio signal at a given location. If the "output" of the model is a probability distribution of the signal's properties, the model is *statistical*, as opposite to a *deterministic* model which gives a single value for each of the predicted properties. Signal strength[1], denoted by s, will be used throughout this chapter, although the approach is applicable to any observable property or properties of the signal.

[1] In principle the terms *field strength* and *signal strength* refer to the magnitude of an electric field. However, in this context they are used as if they were synonymous to the received *power* which depends on the magnitude of the electric field through the power density flow and the receiving antenna as described in Chapter 2.

4.1.1 Single Transmitter Model

The log-loss model defined by Equation (2.9) in Chapter 2 can be used as a statistical propagation model as long as the distribution of the error term e is defined. If a zero-mean Gaussian distribution with a constant variance is used, the model is a linear regression model[2] with three parameters: two regression coefficients, β_0 and β_1, which define the mean value of the signal strength at a given distance, and the variance of e denoted by σ^2. The mean value of the signal strength is given by

$$\mu(d, p, \theta) = p + \beta_0 + \beta_1 \log(d), \tag{4.1}$$

where p is the transmitted power in decibels, and θ denotes the set of parameters.

The transmitters of cellular networks are often directed to some *direction of transmission* to which the transmitted power is higher than to other directions. Therefore, the log-loss model can be improved by adding a term which depends on the deviation between the direction of the receiver and the direction of transmission. The deviation is denoted by δ, and its values are between zero and 180 degrees (see Figure 4.1).

In addition to the parameters of the log-loss model, the improved log-loss model has an additional parameter, β_2, which is related to δ. The mean value of s is given by

$$\begin{aligned} \mu(d, \delta, p, \theta) &= p + \beta_0 + \beta_1 \log(d) + \beta_2 \delta \log(d) \\ &= p + \beta_0 + (\beta_1 + \beta_2 \delta) \log(d). \end{aligned} \tag{4.2}$$

It can be seen on the second row of Equation (4.2) that if the deviation, δ, is constant, the improved model is identical to the normal log-loss model with β_1 replaced by $\beta_1 + \beta_2 \delta$. In other words, attenuation obeys the log-loss model along any straight line originating from the transmitter. Figure 4.2 shows values of μ evaluated using Equation (4.2).

The distribution of s is Gaussian with the following p.d.f.:

$$f(s \mid d, \delta, p, \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2} \left(\frac{s - \mu(d, \delta, p, \theta)}{\sigma} \right)^2 \right]. \tag{4.3}$$

[2]It is important to remember that the term *linear* does not imply that the model can not contain non-linear functions. For instance, in our model the average signal strength is linear with respect to the *logarithm* of the distance.

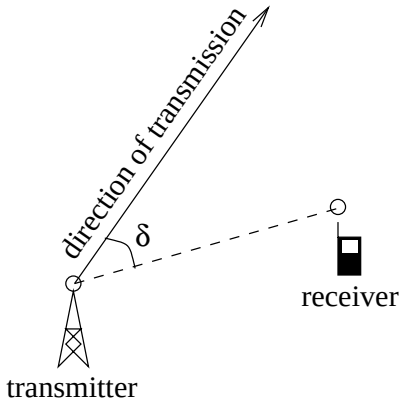

Figure 4.1: The deviation, δ, between the direction of transmission and the direction of the receiver as measured from the transmitter.

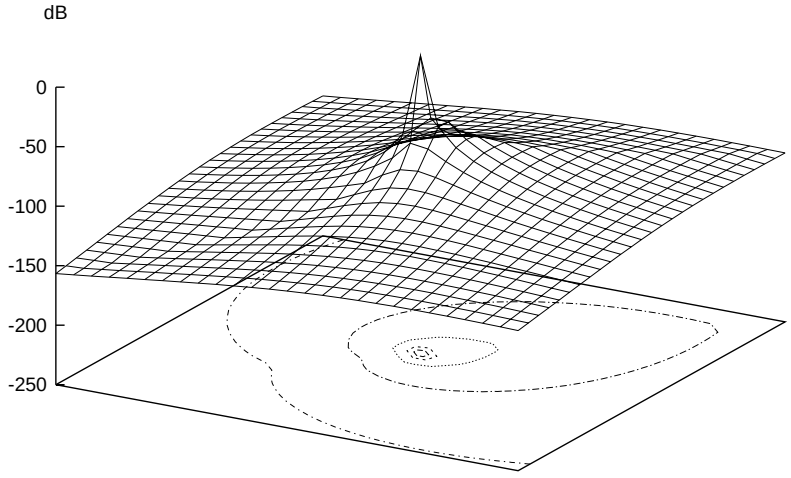

Figure 4.2: An illustration of the average attenuation evaluated using Equation (4.2). The transmitter is located in the center of the area and its direction of transmission is towards the upper right corner of the area.

4.1.2 Multiple Transmitters Model

We have now described how the distribution of the signal strength is evaluated with respect to one transmitter. Let us now extend the model to deal with several transmitters. First, because several *channels*, each operating on a separate frequency range, are used simultaneously in cellular networks, there are actually as many signal strength variables as there are channels. Let s_j denote the signal strength of channel j, and c_i denote the channel of transmitter i. Second, transmitters are classified depending on their transmission properties and location with respect to buildings. For instance, the signal received from an indoor transmitter is usually weaker than the signal received from an outdoor transmitter at the same distance, because of the attenuation caused by buildings. In order to take these differences into account, we use different parameters for each transmitter type. Let $t_i \in \{1, \ldots, k\}$, where k is the number of different types, denote the type of transmitter i, and let $\boldsymbol{\theta}$ denote the parameters of all transmitter types. The parameters of transmitter type j are denoted by $\theta(j)$. Thus, the parameters of transmitter i are denoted by $\theta(t_i)$.

If there are two transmitters on the same channel they cause interference and it is difficult to predict the resulting field strength. However, the situations in which two transmitters on a same channel are close to each other are intentionally avoided while planning the layout of the network, and thus the signal strength of no more than one transmitter is likely to be significant. In such cases we assume the signal strength at a given location to be distributed as if the only transmitter were the one whose mean signal strength according to Equation (4.2) is higher at that particular location. The strongest signal is not necessarily the one of the nearest transmitter, because of the effect of the direction of transmission and differences in the parameters between different transmitter types.

Thus, each transmitter has location, denoted by l_i, type, denoted by t_i, direction of transmission, denoted by α_i, and transmitted power, denoted by p_i. Let g_j denote the p.d.f of the signal strength of channel j, given that the measurement is performed at location l. It is given by the equation

$$g_j(s \mid l, \boldsymbol{\theta}) \stackrel{\text{df.}}{=} f\big(s \mid d(l, l_i),\ \Delta(l, l_i, \alpha_i),\ p_i,\ \theta(t_i)\big) \tag{4.4}$$

where $d(l, l_i)$ is the distance between locations l and l_i, $\Delta(l, l_i, \alpha_i)$ is the deviation at location l with respect to a transmitter located at l_i and directed to α_i. The index i is chosen so that it maximizes the mean signal strength:

$$i = \underset{\{\, i \,:\, c_i = j \,\}}{\operatorname{argmax}} \mu\big(d(l, l_i),\ \Delta(l, l_i, \alpha_i),\ p_i,\ \theta(t_i)\big), \tag{4.5}$$

where function μ is given by Equation (4.2). Thus, when the propagation parameters, and the location, channel, direction of transmission, and transmitted power of the transmitters are fixed, an estimate of the distribution of s_j, for each channel j, is available for every location. We shall next consider how to deal with the unknown propagation parameters.

4.2 Estimation of Propagation Parameters

In most propagation models there are some parameters whose values can not be derived from the underlying theory. These parameters are typically somehow related to the environment and hence, there are no universally good values for them. In such cases, it is obligatory to use empirical data to obtain information about the parameter values. Note, however, that it is generally unjustified to assume the existence of some *true* parameter values which are referred to in the following quote:

> " In many statistics problems, the probability distribution that generated the experimental data is completely known expect for the values of one or more parameters. " (DeGroot, 1986, p. 311)

When modeling phenomena as complex as radio wave propagation the assumption is certainly incorrect. Instead of trying to find the real parameter values, a more realistic goal would be to maximize the expected predictive accuracy. Because even this is often beyond our capabilities, it is a common practice to maximize the *likelihood* of the parameters. This results in the simple, but generally unjustified and suboptimal maximum likelihood approach, which we will use.

In our case the propagation parameters for each transmitter type are estimated from data consisting of signal strength measurements, each labeled with the corresponding channel and location of the receiver. The transmitter information consists of the already mentioned properties, namely the location, channel, direction of transmission, and transmitter power of each transmitter. Based on the data we need to estimate the parameters $\beta_0, \beta_1, \beta_2$ and σ of Equations (4.2) and (4.3) for each transmitter type.

As a preprocessing step the transmitter information is combined with the signal strength measurements in order to produce a table consisting of the following columns:

1. received signal strength, denoted by $\mathbf{s} = s^{(1)}, \ldots, s^{(n)}$,
2. transmitter–receiver distance, denoted by $\mathbf{d} = d^{(1)}, \ldots, d^{(n)}$,

3. deviation between the direction of transmission and the direction of the receiver, $\boldsymbol{\delta} = \delta^{(1)}, \ldots, \delta^{(n)}$, and
4. transmitted power, $\mathbf{p} = p^{(1)}, \ldots, p^{(n)}$.
5. transmitter type, denoted by $\mathbf{t} = t^{(1)}, \ldots, t^{(n)}$,

Filling in fields 2–5 requires that the source of each measured signal is identified, also in cases where there are several transmitters on the same channel. In such cases we assume that the signal is coming from the transmitter which is nearest to the receiver, although in principle Equation (4.5) should be used, and these two criteria do not always agree. This is a deliberate pragmatic choice: Using Equation (4.5) would require treating the ambiguous cases as missing data because $\boldsymbol{\theta}$, whose value is unknown, appears in the equation.

We shall next describe how to obtain maximum likelihood estimates (MLEs) of the parameters, or approximations thereof, from empirical data. First, the simple case where none of the data is missing is discussed, after which a solution to the realistic missing data case is presented.

4.2.1 Maximum Likelihood from Complete Data

Evaluating the MLEs from complete data can be performed easily by exploiting the fact that, what we have is in effect a set of linear regression models. By minor rearrangements, the problem can be formulated in such a way that standard methods for solving MLEs for linear regression models can be applied.

Given fully observed data vectors, $\mathbf{s}$, $\mathbf{d}$, $\boldsymbol{\delta}$, $\mathbf{p}$ and $\mathbf{t}$, the likelihood, $\mathcal{L}(\boldsymbol{\theta})$, is a product of the conditional p.d.f.'s of the individual observations:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f\left(s^{(i)} \mid d^{(i)}, \delta^{(i)}, p^{(i)}, \theta(t^{(i)})\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma(t^{(i)})} \exp\left(-\frac{1}{2}\left(\frac{s^{(i)} - \mu^{(i)}}{\sigma(t^{(i)})}\right)^2\right), \end{aligned} \tag{4.6}$$

where $\sigma(t^{(i)})$ denotes the parameter σ of transmitter type $t^{(i)}$, and $\mu^{(i)}$ is given by

$$\mu^{(i)} \stackrel{\text{df.}}{=} \mu\left(d^{(i)}, \delta^{(i)}, p^{(i)}, \theta(t^{(i)})\right). \tag{4.7}$$

The factorization of $\mathcal{L}(\boldsymbol{\theta})$ is based on the assumption that the variables $s^{(1)}, \ldots, s^{(n)}$ are independent and identically distributed.

The terms in (4.6) can be arranged in k groups where each group corresponds to one of the k transmitter types. Within each group j we use the notation $s^{\langle i,j\rangle}$, $d^{\langle i,j\rangle}$, $\delta^{\langle i,j\rangle}$, $p^{\langle i,j\rangle}$, and $\mu^{\langle i,j\rangle}$ to denote the observations related to the group. For instance, $s^{\langle 1,2\rangle}$ denotes the signal strength of the first observation having $t^{(i)} = 2$. The value of $t^{(i)}$, and hence also $\theta(t^{(i)})$ and $\sigma(t^{(i)})$, is constant for each i belonging to the same group, and we can write the likelihood function as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{j=1}^{k}\left[\prod_{i=1}^{n(j)}\frac{1}{\sqrt{2\pi}\,\sigma(j)}\,\exp\left(-\frac{1}{2}\left(\frac{s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}}{\sigma(j)}\right)^2\right)\right] \\ &= \prod_{j=1}^{k}\left[\left(\frac{1}{\sqrt{2\pi}\,\sigma(j)}\right)^{n(j)}\,\exp\left(-\frac{\mathrm{SSE}(j)}{2\,\sigma(j)^2}\right)\right], \qquad (4.8)\end{aligned}$$

where $n(j)$ is the number of terms having $t^{(i)} = j$, and the sum of squared errors, $\mathrm{SSE}(j)$, is given by

$$\begin{aligned}\mathrm{SSE}(j) &\stackrel{\text{df.}}{=} \sum_{i=1}^{n(j)}\left(s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2 \\ &= \sum_{i=1}^{n(j)}\Big(s^{\langle i,j\rangle} - p^{\langle i,j\rangle} - \beta_0(j) \\ &\qquad - \beta_1(j)\log\left(d^{\langle i,j\rangle}\right) - \beta_2(j)\,\delta^{\langle i,j\rangle}\log\left(d^{\langle i,j\rangle}\right)\Big)^2. \qquad (4.9)\end{aligned}$$

It is now fairly easy to verify that each of the k terms in (4.8) is determined by the parameters of the corresponding transmitter type only, and we can maximize $\mathcal{L}(\boldsymbol{\theta})$ by maximizing each term at a time. Those familiar with linear regression notice that the terms are actually the ones used for obtaining MLEs for linear regression models.

Without proof—one can be found in (DeGroot, 1986)—we state that the maximum likelihood estimates[3] $\widehat{\beta_0(j)}$, $\widehat{\beta_1(j)}$ and $\widehat{\beta_2(j)}$ are independent of $\widehat{\sigma(j)}$, and that they can be obtained by minimizing $\mathrm{SSE}(j)$. Using matrix notation[4] the solution is given by

$$\widehat{\boldsymbol{\beta}(j)} = \left(\mathbf{Z}(j)^T\,\mathbf{Z}(j)\right)^{-1}\mathbf{Z}(j)^T\,\mathbf{Y}(j), \qquad (4.10)$$

[3] We denote the MLE of a variable by $\widehat{X}$.

[4] The notation $\mathbf{A}^T$ denotes the transpose of matrix $\mathbf{A}$, and the inverse of matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{-1}$.

where $\widehat{\boldsymbol{\beta}(j)}$, $\mathbf{Z}(j)$, and $\mathbf{Y}(j)$ are defined as

$$\widehat{\boldsymbol{\beta}(j)} \stackrel{\text{df.}}{=} \begin{bmatrix} \widehat{\beta_0(j)} \\ \widehat{\beta_1(j)} \\ \widehat{\beta_2(j)} \end{bmatrix}, \quad \mathbf{Y}(j) \stackrel{\text{df.}}{=} \begin{bmatrix} s^{\langle 1,j \rangle} - p^{\langle 1,j \rangle} \\ s^{\langle 2,j \rangle} - p^{\langle 2,j \rangle} \\ \vdots \\ s^{\langle n(j),j \rangle} - p^{\langle n(j),j \rangle} \end{bmatrix},$$

$$\mathbf{Z}(j) \stackrel{\text{df.}}{=} \begin{bmatrix} 1 & \log(d^{\langle 1,j \rangle}) & \delta^{\langle 1,j \rangle} \log(d^{\langle 1,j \rangle}) \\ 1 & \log(d^{\langle 2,j \rangle}) & \delta^{\langle 2,j \rangle} \log(d^{\langle 2,j \rangle}) \\ \vdots & \vdots & \vdots \\ 1 & \log(d^{\langle n(j),j \rangle}) & \delta^{\langle n(j),j \rangle} \log(d^{\langle n(j),j \rangle}) \end{bmatrix}. \tag{4.11}$$

Finally, the MLE of $\sigma(j)$ can be obtained from

$$\widehat{\sigma(j)} = \sqrt{\frac{\text{SSE}(j)}{n(j)}}. \tag{4.12}$$

The value of $\text{SSE}(j)$ is obtained by fixing the values of the β-parameters to their MLEs given by Equation (4.10). Equations (4.10) and (4.12) give us the MLEs of the parameter in closed form when the data is complete. The somewhat more complicated missing data case is discussed in the next section.

Example 1. *Figure 4.3 shows an artificial data set containing 66 observations. The path loss values plotted on the vertical axis are the same values that are contained in matrix* $\mathbf{Y}(j)$*. The data was generated by sampling from the propagation model presented in this chapter. Table 4.1 shows the parameters used for generating the data, and the MLEs evaluated using Equations* (4.10) *and* (4.12).

parameter	*actual*	*MLE*
β_0	-30.00	-36.43
β_1	-10.00	-8.77
β_2	-0.0400	-0.0413
σ	10.0	9.8

Table 4.1: The actual parameter values used when generating the data set of Example 1, and the corresponding MLEs.

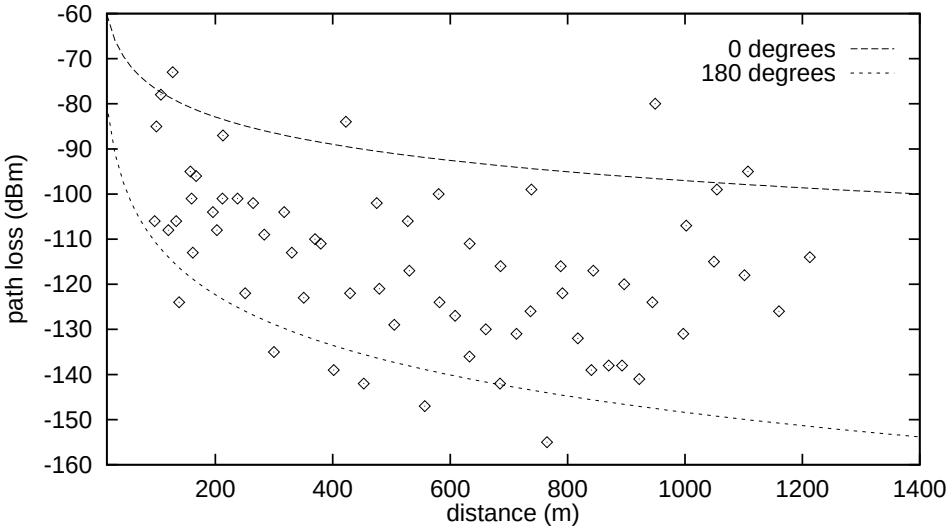

Figure 4.3: Mean path loss curves obtained from sample data. Small blots represent observed path loss values at varying distances from the transmitter. The two curves show the mean path loss to the direction of transmission ($\delta = 0°$), and to the opposite direction ($\delta = 180°$).

4.2.2 Maximum Likelihood from Incomplete Data

In our case some of the signal strengths can not be directly observed because of two reasons related to the measuring device. First, the signal strength values are *binned*, i.e. rounded to the accuracy of one decibelmilliwatt. Second, after each measurement operation the signal strength of only seven channels—those with the strongest signal—is reported. The only information about the other channels is that their signal strength value does not exceed any of the seven known values. In such cases we say that the signal strength variable is *truncated* at a point given by the smallest of the seven known values.

Let the random vector $\mathbf{o} = o^{(1)}, \ldots, o^{(n)}$ denote the observations. For simplicity we assume that the observations are labeled in such a way that the first m variables correspond to binned observations and the $n - m$ other ones correspond to truncated observations. Thus, the relationship between $\mathbf{o}$ and $\mathbf{s}$ is defined by

$$\begin{aligned} o^{(i)} - \frac{\epsilon}{2} \leq\ & s^{(i)} < o^{(i)} + \frac{\epsilon}{2} \quad \text{for } i \in \{1, \ldots, m\} \\ & s^{(i)} \leq o^{(i)} + \frac{\epsilon}{2} \quad \text{for } i \in \{m+1, \ldots, n\}, \end{aligned} \tag{4.13}$$

where the accuracy is determined by ϵ, whose value can be, for instance, 1.0 dBm.

The likelihood function for incomplete data, $\mathcal{L}_{\mathcal{I}}$ (the $\mathcal{I}$ stands for *incomplete*) for an observation vector $\mathbf{o}$ is given by

$$\mathcal{L}_{\mathcal{I}}(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int_{o^{(i)}-\frac{\epsilon}{2}}^{o^{(i)}+\frac{\epsilon}{2}} f\big(s \mid d^{(i)}, \delta^{(i)}, p^{(i)}, \theta(t^{(i)})\big)\, ds \prod_{i=m+1}^{n} \int_{-\infty}^{o^{(i)}+\frac{\epsilon}{2}} f\big(s \mid d^{(i)}, \delta^{(i)}, p^{(i)}, \theta(t^{(i)})\big)\, ds. \tag{4.14}$$

The equation is analogous to the likelihood function for complete data given by Equation (4.6). However, it is not straightforward to derive a closed form solution analogous to the complete-data solution.[5] Instead, there is a method which can be used to approximate a local maximum of the likelihood function from incomplete data, namely the Expectation–Maximization (EM) algorithm (Dempster *et al.* , 1977; McLachlan & Krishnan, 1997).

The EM algorithm can be applied whenever it is possible to evaluate the expected value of the logarithm of the complete data likelihood (log-likelihood). In order to evaluate the expected log-likelihood we need a probability distribution for the missing signal strength values. In the EM algorithm the distribution is obtained by fixing the parameters to some hypothetical values, say $\boldsymbol{\theta}^{(r)}$. The expectation of the log-likelihood function, denoted by $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$, is then evaluated in the *expectation step* using the equation

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) \stackrel{\text{df.}}{=} E\big\{\ln \mathcal{L}(\boldsymbol{\theta}) \mid \boldsymbol{\theta}^{(r)}\big\}, \tag{4.15}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is the complete-data likelihood, given by Equation (4.6). In the *maximization step* the parameter values are replaced by ones which maximize the expected log-likelihood, thus giving

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}), \tag{4.16}$$

where $\boldsymbol{\theta}^{(r)}$ denotes the parameters on step r. The algorithm consists of repeating these two steps, one after the other. It can be shown that the likelihood of the parameter values is never decreased during an iteration. Thus, if the algorithm converges, it converges to a local maximum of the likelihood function.

[5] As usual, the phrase "not straightforward" is used as an indirect way to say that the author is not aware of such a solution.

It now remains to be shown how to obtain a set of parameter values from Equation (4.16). By taking the logarithm of $\mathcal{L}(\boldsymbol{\theta})$, given by Equation (4.6), and substituting it into Equation (4.15) we get

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = E\left\{ \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \ln\big(\sigma(t^{(i)})\big) - \frac{1}{2}\left(\frac{s^{(i)} - \mu^{(i)}}{\sigma(t^{(i)})} \right)^2 \right] \;\middle|\; \boldsymbol{\theta}' \right\}. \tag{4.17}$$

By switching the order of the expectation and sum operators and taking terms that are independent of $\mathbf{s}$ outside of the expectation, the equation becomes

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \ln\big(\sigma(t^{(i)})\big) - \frac{1}{2\,\sigma(t^{(i)})^2} E\left\{ \big(s^{(i)} - \mu^{(i)}\big)^2 \;\middle|\; \boldsymbol{\theta}' \right\} \right]. \tag{4.18}$$

Like in the complete data case the function to be optimized, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$, can be rearranged into to k terms each depending on the parameters of one transmitter type only. Therefore we can find the maximizing parameters for each transmitter type at a time. Let $Q_j(\theta(j), \theta'(j))$ denote the terms of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that are determined by the parameters of transmitter type j. For each type j, we have

$$Q_j(\theta(j), \theta'(j)) = -\frac{n(j)}{2}\ln(2\pi) - n(j)\ln(\sigma(j)) - \frac{1}{2\,\sigma(j)^2}\,\mathrm{SESE}(j), \tag{4.19}$$

where $\mathrm{SESE}(j)$ is the sum of expected squared errors given by

$$\mathrm{SESE}(j) \stackrel{\text{df.}}{=} \sum_{i=1}^{n(j)} \left[E\left\{ \big(s^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\big)^2 \;\middle|\; \theta'(j) \right\} \right]. \tag{4.20}$$

The value of $\boldsymbol{\beta}(j)$ which maximizes Equation (4.19) is given by (see Appendix A for a proof)

$$\widehat{\boldsymbol{\beta}(j)} = \big(\mathbf{Z}(j)^T\,\mathbf{Z}(j)\big)^{-1}\,\mathbf{Z}(j)^T\,\mathbf{Y}(j) \tag{4.21}$$

where $\widehat{\boldsymbol{\beta}(j)}$,[6] and $\mathbf{Z}(j)$ are defined by Equation (4.11), and $\mathbf{Y}(j)$ is defined

[6]The notation $\widehat{\boldsymbol{\beta}(j)}$ is used, although the solution is in fact the maximizer of $Q_j(\theta(j), \theta'(j))$, not the (incomplete-data) likelihood.

as

$$\mathbf{Y}(j) \stackrel{\text{df.}}{=} \begin{bmatrix} E\{s^{\langle 1,j\rangle} \mid \theta'(j)\} - p^{\langle i,j\rangle} \\ E\{s^{\langle 2,j\rangle} \mid \theta'(j)\} - p^{\langle 2,j\rangle} \\ \vdots \\ E\{s^{\langle n(j),j\rangle} \mid \theta'(j)\} - p^{\langle n(j),j\rangle} \end{bmatrix}. \tag{4.22}$$

Thus, in order to obtain estimates of the β-parameters we need to evaluate the expected value of $s^{\langle i,j\rangle}$, for each $j \in \{1,\ldots,k\}$ and $i \in \{1,\ldots,n(j)\}$. For binned observations, the expected value of $s^{\langle i,j\rangle}$ is (see Appendix A for a proof)

$$E\{s^{\langle i,j\rangle} \mid \theta'(j)\} = \frac{\left(\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - \exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)\sigma'(j)}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} + \mu'^{\langle i,j\rangle}, \tag{4.23}$$

where Φ is the cumulative distribution function of a Gaussian distribution with zero mean and unity variance; $\mu'^{\langle i,j\rangle}$ is the mean signal strength value according to the log-loss model with parameters $\theta'(j)$:

$$\mu'^{\langle i,j\rangle} \stackrel{\text{df.}}{=} \mu\big(d^{\langle i,j\rangle}, \delta^{\langle i,j\rangle}, p^{\langle i,j\rangle}, \theta'(j)\big); \tag{4.24}$$

and $a^{\langle i,j\rangle}$ and $b^{\langle i,j\rangle}$ are given by

$$a^{\langle i,j\rangle} \stackrel{\text{df.}}{=} \frac{o^{\langle i,j\rangle} - \frac{\epsilon}{2} - \mu'^{\langle i,j\rangle}}{\sigma'(j)}, \qquad b^{\langle i,j\rangle} \stackrel{\text{df.}}{=} \frac{o^{\langle i,j\rangle} + \frac{\epsilon}{2} - \mu'^{\langle i,j\rangle}}{\sigma'(j)}. \tag{4.25}$$

Because the value of $s^{\langle i,j\rangle}$ is known to be within the range $o^{\langle i,j\rangle} \pm \frac{\epsilon}{2}$, its expected value must also be within the same range. The difference between the exact solution and $o^{\langle i,j\rangle}$ is bound by the equation

$$\left|E\{s^{\langle i,j\rangle} \mid \theta'(j)\} - o^{\langle i,j\rangle}\right| \leq \frac{\epsilon}{2}. \tag{4.26}$$

Thus, the expectation can be approximated by $o^{\langle i,j\rangle}$.

For truncated observations, the expectation of $s^{\langle i,j\rangle}$ is given by (see Appendix A for a proof)

$$E\{s^{\langle i,j\rangle} \mid \theta'(j)\} = -\frac{\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\,\sigma'(j)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + \mu'^{\langle i,j\rangle}, \tag{4.27}$$

where $b^{\langle i,j\rangle}$ is given by Equation (4.25).

The value of $\sigma(j)$ maximizing (4.19) is given by (see Appendix A for a proof)

$$\widehat{\sigma(j)} = \sqrt{\frac{\mathrm{SESE}(j)}{n(j)}}. \tag{4.28}$$

In order to evaluate $\mathrm{SESE}(j)$, which appears in Equation (4.28), we need a closed form solution for the expected squared error $E\left\{(s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle})^2 \mid \theta'(j)\right\}$. For binned observations, it is given by (see Appendix A for a proof)

$$\begin{aligned}
&E\left\{\left(s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2 \,\middle|\, \theta'(j)\right\} \\
&= \frac{\sigma'(j)^2\left(a^{\langle i,j\rangle}\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)}{\sqrt{2\pi}\left(\Phi(b^{\langle i,j\rangle})-\Phi(a^{\langle i,j\rangle})\right)} + \sigma'(j)^2 \\
&+ \frac{2\,\sigma'(j)\left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)\left(\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - \exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)}{\sqrt{2\pi}\left(\Phi(b^{\langle i,j\rangle})-\Phi(a^{\langle i,j\rangle})\right)} \\
&+ \left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2,
\end{aligned} \tag{4.29}$$

where $a^{\langle i,j\rangle}$, and $b^{\langle i,j\rangle}$ are given by Equation (4.25), and $\mu^{\langle i,j\rangle}$ is obtained by using the estimates of the β-parameters given by Equation (4.21). A reasonable approximation to Equation (4.29) is given by $(o^{\langle i,j\rangle}-\mu^{\langle i,j\rangle})^2$, because $s^{\langle i,j\rangle}$ is known to be within the range $o^{\langle i,j\rangle} \pm \frac{\epsilon}{2}$.[7]

For truncated observations, the expected squared error is given by (see Appendix A for a proof)

$$\begin{aligned}
&E\left\{\left(s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2 \,\middle|\, \theta'(j)\right\} \\
&= -\frac{\sigma'(j)^2\, b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + \sigma'(j)^2 \\
&- \frac{2\,\sigma'(j)\left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + \left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2.
\end{aligned} \tag{4.30}$$

By looking at Equation (4.30) one can see that the last two terms can be ignored, if we assume that the difference $|\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}|$, i.e. the difference between two consequent estimates of the mean signal strength value, is very small. Unless the EM-algorithm does not converge at all, this is guaranteed to be the case in the long run.

[7] Such an approximation is in fact implicitly used every time finite-precision values are treated as precise ones. This was the case also in the complete-data case of the previous section.

We now have closed form solutions for the parameters maximizing Equation (4.19): Equation (4.21) for $\boldsymbol{\beta}(j)$ and Equation (4.28) for $\sigma(j)$. By using them for each transmitter type $j \in \{1, \ldots, k\}$ we obtain the parameter vector $\boldsymbol{\theta}$ maximizing Equation (4.15). This vector is all that is needed to solve Equation (4.16), and in fact, all that is needed to perform an iteration of the EM algorithm.

Example 2. *Figure 4.4 shows an artificial data set containing 66 observations, 37 of which are binned, while the 29 other ones are truncated. For truncated observations, the figure shows the truncation point which is known to be higher than the unknown path loss value. Table 4.2 shows the parameters used for generating the data, and estimates obtained with the EM-algorithm. The data set of Example 2 is the same as the one used in Example 1 with the exception that in Example 2 some of the observations are truncated. Note that the parameter estimates given in Tables 4.1 and 4.2 are very much alike, which shows that the difference between the actual values and the ones obtained with the EM-algorithm are mainly caused by the relatively small sample size, not for instance, by the incomplete data.*

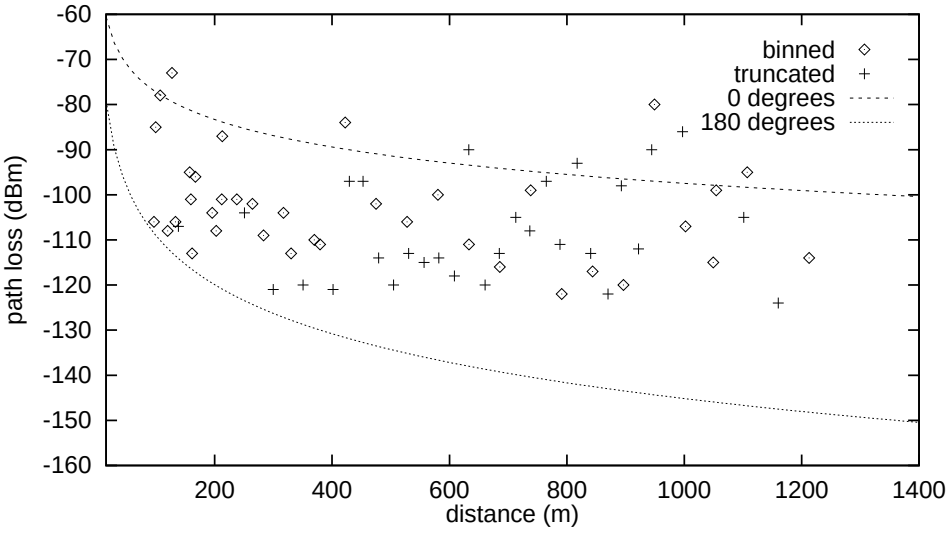

Figure 4.4: Mean path loss curves obtained from sample data. Small symbols represent binned ($\diamond$) and truncated ($+$) path loss values at varying distances from the transmitter. The two curves show the mean path loss to the direction of transmission ($\delta = 0°$), and to the opposite direction ($\delta = 180°$).

iteration	β_0	β_1	β_2	σ
1	-51.84	-6.38	-0.0280	9.4
2	-48.57	-6.98	-0.0313	9.4
3	-44.58	-7.59	-0.0337	9.5
4	-42.01	-7.98	-0.0352	9.5
5	-40.36	-8.23	-0.0362	9.6
6	-39.27	-8.40	-0.0368	9.6
7	-38.52	-8.51	-0.0373	9.7
8	-38.01	-8.59	-0.0376	9.7
9	-37.65	-8.65	-0.0378	9.7
10	-37.40	-8.69	-0.0379	9.7
11	-37.22	-8.71	-0.0380	9.8
12	-37.09	-8.73	-0.0381	9.8
13	-37.00	-8.75	-0.0382	9.8
14	-36.94	-8.76	-0.0382	9.8
15	-36.89	-8.76	-0.0383	9.8
16	-36.85	-8.77	-0.0383	9.8
17	-36.83	-8.77	-0.0383	9.8
18	-36.81	-8.77	-0.0383	9.8
19	-36.80	-8.78	-0.0383	9.8
20	-36.79	-8.78	-0.0383	9.8
21	-36.79	-8.78	-0.0383	9.8
22	-36.78	-8.78	-0.0383	9.8
23	-36.78	-8.78	-0.0383	9.8
24	-36.78	-8.78	-0.0383	9.8
25	**-36.77**	**-8.78**	**-0.0383**	**9.8**
actual	-30.00	-10.00	-0.0400	10.00

Table 4.2: The values of the parameter estimates for EM-iterations 1–25 with the data set of Example 2. The algorithm has converged with the precision used in the table by iteration 25. The actual parameter values used when generating the data set are shown at the bottom of the table.

4.3 Estimation of Location

Given the estimates of the propagation parameters $\hat{\boldsymbol{\theta}}$, the p.d.f. of the field strength of channel j at location l is given by $g_j(s_j \mid l, \hat{\boldsymbol{\theta}})$, where g_j is defined by Equation (4.4). The posterior p.d.f. of the location variable l is given by the Bayes rule[8]:

$$p(l \mid \mathbf{s}, \hat{\boldsymbol{\theta}}) = \frac{g(\mathbf{s} \mid l, \hat{\boldsymbol{\theta}})\, \pi(l)}{\int g(\mathbf{s} \mid l', \hat{\boldsymbol{\theta}})\, \pi(l')\, dl'}, \tag{4.31}$$

where $\mathbf{s}$ is a vector consisting of the field strength values s_j for each channel j, and $g(\mathbf{s} \mid l, \hat{\boldsymbol{\theta}})$ is the likelihood function given by

$$g(\mathbf{s} \mid l, \hat{\boldsymbol{\theta}}) = \prod_j g_j(s_j \mid l, \hat{\boldsymbol{\theta}}), \tag{4.32}$$

and π is the prior p.d.f. of the location variable.

However, Equation (4.32) is not directly applicable for practical location estimation purposes if some of the signal strength observations are truncated.[9] It is not the actual signal strength vector, $\mathbf{s}$, that is observed, but the observation vector, $\mathbf{o}$, whose relation to $\mathbf{s}$ is the following:

$$\begin{aligned} o_j - \frac{\epsilon}{2} \leq\ & s_j < o_j + \frac{\epsilon}{2} \quad \text{if } j \in \mathcal{B} \\ & s_j \leq o_j + \frac{\epsilon}{2} \quad \text{if } j \in \mathcal{T}, \end{aligned} \tag{4.33}$$

where $\mathcal{B}$ is the set of binned channels, and $\mathcal{T}$ is the set of truncated channels, and the accuracy of the measurements is determined by ϵ.

Now that the propagation parameters, $\hat{\boldsymbol{\theta}}$, are fixed, the likelihood function is defined with respect to the location variable, l, and thus, the likelihood function is given by

$$\begin{aligned} g(\mathbf{o} \mid l, \hat{\boldsymbol{\theta}}) = & \prod_{j \in \mathcal{B}} \int_{o_j - \frac{\epsilon}{2}}^{o_j + \frac{\epsilon}{2}} g_j\big(s \mid l, \hat{\boldsymbol{\theta}}\big)\, ds \\ & \prod_{j \in \mathcal{T}} \int_{-\infty}^{o_j + \frac{\epsilon}{2}} g_j\big(s \mid l, \hat{\boldsymbol{\theta}}\big)\, ds. \end{aligned} \tag{4.34}$$

[8] The application of the Bayes rule might be opposed by some people who prefer traditional statistical theory over its Bayesian correspondent (Box & Tiao, 1973; Berger, 1980). The primary concern of the opponents is usually related to the concept of prior distributions. However, in this case the results obtained with traditional statistical methods would be similar to the ones presented here, as we will note later.

[9] If there are no truncated observations, i.e. all the observations are binned, Equation (4.32) is applicable because one can use the center points of the bins as approximations to the actual values of the signal strength variables, unless the bins are very wide.

The corresponding posterior p.d.f. of the location variable is then

$$p(l \mid \mathbf{o}, \hat{\boldsymbol{\theta}}) = \frac{g(\mathbf{o} \mid l, \hat{\boldsymbol{\theta}})\, \pi(l)}{\int g(\mathbf{o} \mid l', \hat{\boldsymbol{\theta}})\, \pi(l')\, dl'}. \tag{4.35}$$

The denominator of the right hand side of Equation (4.35) is constant with respect to l and thus, the posterior p.d.f. of the location variable is proportional to the numerator:

$$p(l \mid \mathbf{o}, \hat{\boldsymbol{\theta}}) \propto g(\mathbf{o} \mid l, \hat{\boldsymbol{\theta}})\, \pi(l). \tag{4.36}$$

In theory, the location variable might be continuous in $\mathbb{R}^2$. In that case, no proper uniform prior π would exist.[10] In practice, however, the location variable is always restricted to some area, and thus, a uniform prior can be used. Of course, if an informative prior is available, it should be used instead.

A location estimate is chosen depending on the penalty function, which defines how different errors are penalized. Two reasonable estimates are the maximum a posteriori location, i.e. the location maximizing Equation (4.36), and the expected value of the location variable. The latter minimizes the expected value of the squared error of the location estimate. If a uniform prior is used, the maximum a posteriori location is the same as the maximum likelihood estimate of l, which would probably be the solution preferred by advocates of the traditional statistical theory.

Because no closed form solution for either the maximum a posteriori value or the expected value is available, the location variable must be discretized. One can, for instance, split the area into squares of fixed size, say 50×50 meters, and use the center point of each square to evaluate the distribution of the field strength variables in that particular square. After discretization the maximum a posteriori value can be obtained simply by going through each of the squares and choosing the value which maximizes Equation (4.36). The expected value of the location variable can be obtained by calculating an average of the location variable weighted by Equation (4.36).

Evaluating the empirical performance of the presented method requires that the layout of a cellular network is known. Such information is only accessible to network operators. We will therefore present only an illustrative example using an artificial network layout, shown in Figure 4.5.

Example 3. *Figure 4.5 represents a hypothetical network layout. Figure 4.6 shows four examples of the posterior p.d.f. of the location variable, the resulting maximum a posteriori location estimate, and the expected value of the*

[10] A prior $\pi(x)$ is proper if it is non-negative, $\pi(x) \geq 0$, for all x, and it integrates to one, $\int \pi(x)\, dx = 1$. A uniform prior $\pi(x) \equiv c$, where c is some constant, violates the latter condition, unless the range of x is finite.

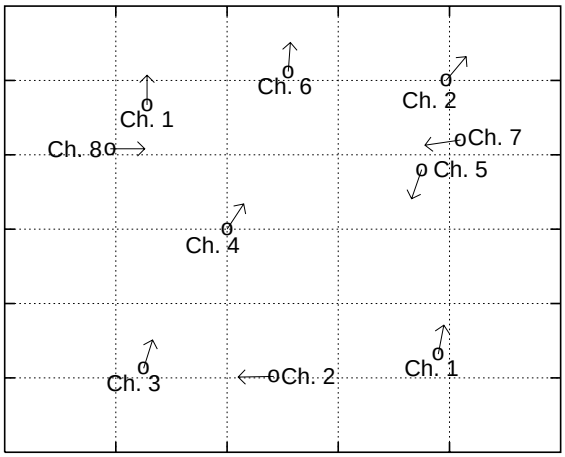

Figure 4.5: A hypothetical network layout consisting of ten transmitters. Arrows indicate direction of transmission, and labels indicate channels. Note that channels one and two are both shared by two transmitters.

location variable with artificial signal strength measurement results. In graph (a), signal strength of channel two is known to be -55 *dBm, and information concerning the other channels is nonexistent. Most of the probability mass is concentrated within two elliptical areas around the channel two transmitters. In graph (b), in addition to channel two, signal strength of channel six is observed to be* -60 *dBm. Observing the signal strength of channel six resolves the ambiguity caused by the fact that two transmitters share channel two, and therefore the p.d.f. in (b) becomes unimodal.*

Graphs (c) and (d) illustrate the effect of truncated observations. In both cases the following signal strength were observed:

Channel 1: -70 dBm	Channel 3: -75 dBm
Channel 4: -70 dBm	Channel 8: -65 dBm

However, the two cases differ with respect to the other channels (2, 5, 6, and 7); in (c) no information concerning them is available, unlike in (d), where the signal strength of those channels is truncated at -75 dBm, i.e. the signal strength values are known to be less than or equal to -75 dBm. By comparing graphs (c) and (d), one can see that the truncated observations can be very useful when estimating location.

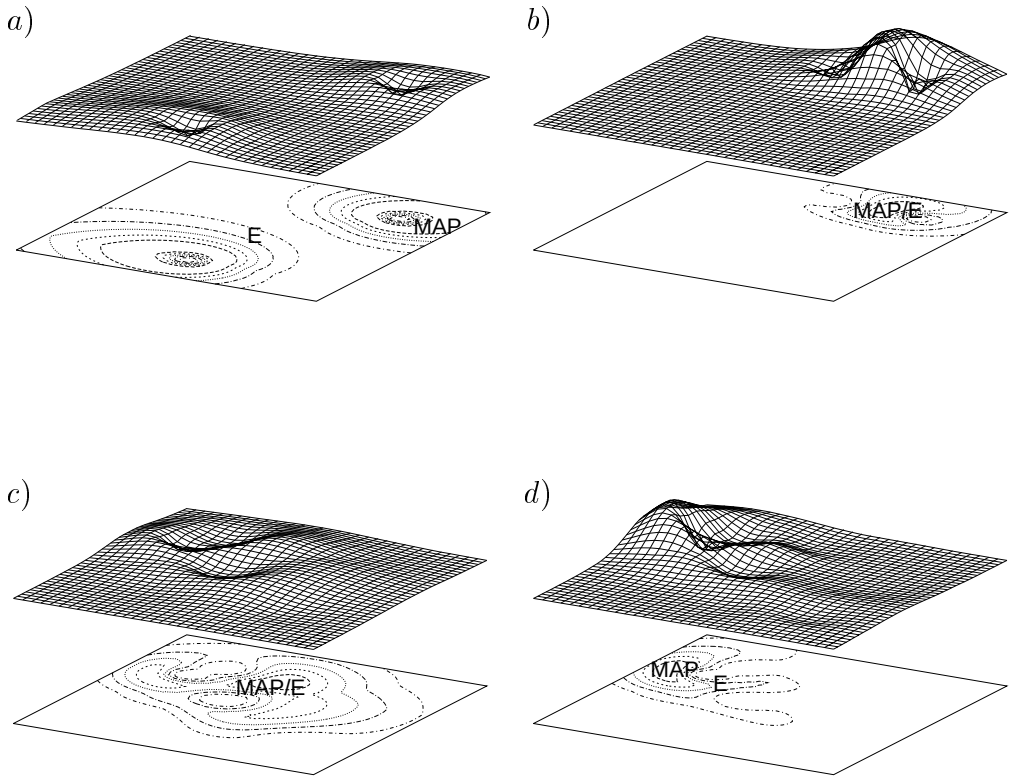

Figure 4.6: Examples of the posterior p.d.f. of the location variable with different signal strength observations. Labels indicate the maximum a posteriori estimate ("MAP"), and the expected value ("E") of the location variable. In graphs (b) and (c) they are practically identical.

Chapter 5

Conclusions

After a short introduction to cellular telephone systems, we presented the principles of radio wave propagation and propagation models. The emphasis was on prediction of signal strength, although several other signal properties can be used in location estimation. An overview of the conventional location estimation methods was then given. In Chapter 4, which is the main contribution of the thesis, a statistical location estimation method based on a propagation prediction model was presented. To conclude, we shall now discuss some lessons learned during the process of developing a location estimation method, and writing this thesis.

The conceptual development of location estimation methods has been modest after the ancient Egyptians and Greeks invented the art of triangulation. The problem has been mostly considered by engineers, and consequently, a majority of proposed solutions are geometric in nature. For instance, the Angle of Arrival method is nothing more than triangulation. In addition to triangulation, some geometric methods, such as Time of Arrival, and Time Difference of Arrival—which is used in the GPS system—are based on distance measurements rather than angle measurements. The geometric solutions work very well in certain situations. However, if the signal propagation environment differs significantly from ideal conditions, the distance or angle measurements are unreliable. In such cases, serious problems occur because the various measurements are inaccurate at best, incompatible at worst. Special *ad hoc* heuristics have to be applied in order to compensate for these errors.

In this thesis, an alternative approach to the location estimation problem was taken, which we call the statistical approach. Here signal properties, such as signal strength, angle of arrival and propagation delay, are treated as random variables which are statistically dependent on the locations of the transmitter and the receiver, and the propagation environment. In this

respect, the conceptual difference between the two approaches is clear: In the geometric approach the reasoning goes from the measured signal properties to the location of the transmitter, whereas in the statistical approach, the emphasis is on the propagation model, which describes the dependency of the measured signal properties on the location variable, that is, the reasoning proceeds from the location to the signal properties. This is the kind of reasoning that is typical to statistics in general. In statistical terms, the propagation model is a sampling distribution whose parameters—in the first phase, the propagation parameters, and in the second phase, the location variable—we wish to estimate.

The problem of incompatible measurements is not present in the statistical approach, unlike the geometric ones, because no matter how unlikely the combination of measurement results, it is always possible. Of course, if the propagation model does not fit well the actual propagation phenomena and the environment, the propagation prediction accuracy, and accordingly, the location estimation accuracy is poor. However, whereas the only possibility to enhance the accuracy of the geometric location estimation methods is to increase the accuracy of the angle and distance measurements, this is not the case with methods based on the statistical approach. Their accuracy can be enhanced also by switching to an other propagation model, which is better suited for predicting the relevant signal properties in the environment in question.

The advantages of the statistical approach include also certain types of flexibility, which presented itself in the present work. In our case, the observations made by the mobile unit in order to be located, were associated with a set of channels whose signal strength was known, and an other set of channels, whose signal strength was only bounded from above. We called the latter kind of partial observations *truncated*. The geometric approach provides no principled way of exploiting the information contained in the truncated observations. However, as we saw in Chapter 4, the statistical approach lends itself easily to exploiting any kind of observations, partial or complete.

To be realistic, one has to say that the few empirical results presented in this thesis suffice to prove nothing more than the *theoretical* validity of the proposed location estimation method. In order to show its practical merits one should provide results obtained with real-world data.

The statistical approach is by no means restricted to the use of signal strength measurements. One could also use angle or timing measurements, as long as the used propagation model is capable of handling them. The flexibility of the approach allows also the fusion of different types of measurement results, for instance, signal strength and timing information. Using

the terminology of Chapter 2, the propagation model considered in this work belongs to the class of general models. In other words, the model does not take into account the effect of the heterogeneity of the propagation environment. Another interesting line of investigation is the application of empirical propagation prediction methods to location estimation. Our guess is that there is some potential in such solutions.

References

Andersen, J.B., Rappaport, T.S., & Yoshida, S. 1995. Propagation Measurements and Models for Wireless Communications Channels. *IEEE Communications Magazine*, **33**(1), 42–49.

Asimov, I. 1966. *Understanding Physics: Light, Magnetism, and Electricity.* New York: NAL Penguin.

ASME, American Society of Mechanical Engineers. 2000. *Quotes on Engineering and Technology.* http://www.asme.org/history/hquote4.html.

Athanasiadou, G.E., Nix, A.R., & McGeehan, J.P. 2000. A Microcellular Ray-Tracing Propagation Model and Evaluation of its Narrow-Band and Wide-Band Predictions. *IEEE Journal on Selected Areas in Communications*, **18**(3), 322–335.

Belle, G., Goldstein, D.B., Humble, R.W., Parker, D.L., O'Brien, C., Matini, A., & Brown, A. 1997 (Sept.). *The U. S. Air Force Academy GPS Flight Experiment Using The Navsys TIDGET.* ION GPS 97. Kansas City.

Berger, J. 1980. *Statistical Decision Theory – Foundations, Concepts, and Methods.* New York: Springer-Verlag.

Box, G.E.P., & Tiao, G.C. 1973. *Bayesian Inference in Statistical Analysis.* Reading: Addison-Wesley.

Bretz, E.A. 2000. X Marks the Spot, Maybe. *IEEE Spectrum*, **37**(4).

Cell-Loc. 2001. *Cellocate Feature Sheet.* www.cell-loc.com.

CellPoint. 2001. *The Cellpoint System.* www.cellpoint.com.

Corazza, G.E., Degli-Esposti, V., Frullone, M., & Riva, G. 1996. A Characterization of Indoor Space and Frequency Diversity by Ray-Tracing Modeling. *IEEE Journal on Selected Areas of Communications*, **14**(3), 411–419.

CPS, Cambridge Positioning Systems. 2001. *How Cursor Works.* www.cursor-system.co.uk.

Damosso, E., & Correia, L.M. (eds). 1998. *Digital Mobile Radio towards Future Generation Systems: COST-231 Final Report.* COST-231.

DeGroot, M.H. 1986. *Probability and Statistics (2nd Ed.).* Reading: Addison-Wesley.

Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the *EM* Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, 1–38.

Driscoll, C. 1998. Wireless Caller Location Systems. *GPS World*, Nov.

ETSI, European Telecommunications Standards Institute. 1998 (Apr.). *Digital Cellular Telecommunications System (Phase 2+); Specification of the SIM Application Toolkit for the Subscriber Identity Module – Mobile Equipment (SIM – ME) Interface.* Technical Specification ETSI-TS-101-267-V6.0.0.

FGCS, Federal Geodetic Control Subcommittee. 2000 (May). *Removal of GPS Selective Availability (SA).* http://www.ngs.noaa.gov/FGCS/info/sans_SA/.

Fleury, B.H., & Leuthold, P.E. 1996. Radiowave Propagation in Mobile Communications: An Overview of European Research. *IEEE Communications Magazine*, **34**(2), 70–81.

FTSC, Federal Telecommunications Standards Committee. 1996 (Aug.). *Glossary of Telecommunication Terms.* Federal Standard 1037C. www.its.bldrdoc.gov/fs-1037/.

Hoppe, R., Wölfle, G., & Landstorfer, F.M. 1999 (Mar.). Fast 3-D Ray Tracing for the Planning of Microcells by Intelligent Preprocessing of the Data Base. *Pages 149–154 of: 3rd European Personal Mobile Communications Conference (EPMCC).*

Kelsey Group. 2000. *Webraska, Alcatel Demo Maps Using Positioning Technology.* kelseygroup.com/clp/clp000202aa.htm.

KSI. 2001. *The Telesentinel System.* www.telesentinel.com.

Latapy, J.-M. 1996. *GSM Mobile Station Locating.* Master's Thesis, Norwegian University of Science and Technology, Trondheim.

McLachlan, G.J., & Krishnan, T. 1997. *The EM Algorithm and Extensions.* New York: John Wiley & Sons.

ModelSoft. 1999 (Dec.). *ModelSoft's GSM Positioning Technology Adopted by Radiolinja.* www.modelsoft.fi/modelsoft/modelsoft/press_02121999.html.

ModelSoft. 2001. *ModelSoft Oy's GSM/PCS Location Technology.* www.modelsoft.fi/modelsoft/roadtracker/.

NAWC, Naval Air Warfare Center. 1997 (Apr.). *Electronic Warfare and Radar Systems Engineering Handbook.* Tech. rept. TP 8347. ewhdbks.mugu.navy.mil/.

Radix Technologies. 2001. *E911 GeoPhone System.* www.radixtek.com/e911.htm.

Rantalainen, T., & Pickford, A. 1999 (Aug.). *E-OTD Description to LCS Stage 2 Annex.* Contribution T1P1.5/99-541. www.t1.org/index/0521.htm.

Rappaport, T.S. 1996. *Wireless Communications: Principles & Practice.* Upper Saddle River: Prentice Hall.

Rappaport, T.S., Reed, J.H., & Woerner, B.D. 1996. Position Location Using Wireless Communications on Highways of the Future. *IEEE Communications Magazine*, Oct., 33–41.

Ruutu, V. 2000. *Network Location.* Lecture Notes of a Mobile Location Seminar, Merito Forum, Helsinki, Finland.

SnapTrack. 2001. *An Introduction to SnapTrack Server-aided GPS Technology.* www.snaptrack.com/atwork.html.

Syrjärinne, J. 2001. *Studies of Modern Techniques for Personal Positioning.* D.Sc. Thesis, Tampere University of Technology.

TruePosition. 2000. *TruePosition Cellular Location System.* www.trueposition.com.

UMTS Forum. 2000 (Sept.). *The UMTS Third Generation Market – Structuring the Service Revenues Opportunities.* Report No. 9.

U.S. Wireless. 1999 (Nov.). *U.S. Wireless Corporation Announces Positive Results of State-audited Montana E-911 Trial.* www.uswcorp.com/USWCMainPages/PressRel/pr45.htm.

U.S. Wireless. 2001. *Location Pattern Matching & The RadioCamera Network.* www.uswcorp.com/USWCMainPages/our.htm.

USAEC, U.S. Army Environmental Center. 1954. *USAEC Transcript of Hearing Before Personnel Security Board in the Matter of J. Robert Oppenheimer.*

Walke, B.H. 1999. *Mobile Radio Networks: Networking and Protocols.* Chichester: John Wiley & Sons.

WAP Forum. 1998 (Apr.). *Wireless Application Protocol: Architecture Specification.*

Willassen, S.Y. 1998 (Dec.). *A Method for Implementing Mobile Station Location in GSM.* Diploma Thesis, Norwegian University of Science and Technology, Trondheim.

Wölfle, G., & Landstorfer, F.M. 1998 (May). Dominant Paths for the Field Strength Prediction. *Pages 552–556 of: 48th IEEE Vehicular Technology Conference (VTC).*

Wölfle, G., & Landstorfer, F.M. 1999. Prediction of the Field Strength inside Buildings with Empirical, Neural, and Ray-optical Prediction Models. *In: 7th COST 259 MCM-Meeting in Thessaloniki.*

Xu, H., Rappaport, T.S., Boyle, R.J., & Schaffner, J.H. 2000. Measurements and Models for 38-GHz Point-to-Multipoint Radiowave Propagation. *IEEE Journal on Selected Areas in Communications*, **18**(3), 310–321.

Appendix A

Proofs

Proof of Equation (4.21)**.** The values of $\beta_0(j)$, $\beta_1(j)$, and $\beta_2(j)$, maximizing $Q_j(\theta(j), \theta'(j))$ can be obtained by setting the partial derivatives of the latter, with respect to the former, to zero. The first two terms of $Q_j(\theta(j), \theta'(j))$, given by Equation (4.19), do not depend on the β-parameters. Thus, the partial derivatives depend on the third term only, and we get

$$\frac{\partial\, Q_j(\theta(j), \theta'(j))}{\partial\, \beta_i(j)} = \frac{\partial\, \frac{1}{2}\, \sigma(j)^{-2}\, \mathrm{SESE}(j)}{\partial\, \beta_i(j)} = \frac{1}{2\, \sigma(j)^2} \frac{\partial\, \mathrm{SESE}(j)}{\partial\, \beta_i(j)}, \quad \text{(A.1)}$$

for each $i \in \{0, 1, 2\}$. Note that $\mathrm{SESE}(j)$ *does* depend on the β-parameters.

In order to obtain the partial derivative of $\mathrm{SESE}(j)$ with respect to the β-parameters, we manipulate $\mathrm{SESE}(j)$. From the definition of $\mathrm{SESE}(j)$ in Equation (4.20) it directly follows that

$$\begin{aligned} \mathrm{SESE}(j) &= \sum_{i=1}^{n(j)} \left[E\left\{ \left(s^{\langle i,j \rangle} - \mu^{\langle i,j \rangle}\right)^2 \;\middle|\; \theta'(j) \right\} \right] \\ &= \sum_{i=1}^{n(j)} \left[E\left\{ \left(s^{\langle i,j \rangle}\right)^2 \;\middle|\; \theta'(j) \right\} - 2\, E\left\{ s^{\langle i,j \rangle}\, \mu^{\langle i,j \rangle} \;\middle|\; \theta'(j) \right\} \right. \\ &\qquad \left. + E\left\{ \left(\mu^{\langle i,j \rangle}\right)^2 \;\middle|\; \theta'(j) \right\} \right]. \quad \text{(A.2)} \end{aligned}$$

Because the value of $\mu^{\langle i,j \rangle}$ does not depend on the unknown variable, $s^{\langle i,j \rangle}$, the former can be taken outside the expectation operators, and we get

$$\mathrm{SESE}(j) = \sum_{i=1}^{n(j)} \left[E\left\{ \left(s^{\langle i,j \rangle}\right)^2 \;\middle|\; \theta'(j) \right\} - 2\, \mu^{\langle i,j \rangle}\, E\left\{ s^{\langle i,j \rangle} \;\middle|\; \theta'(j) \right\} + \left(\mu^{\langle i,j \rangle}\right)^2 \right]. \quad \text{(A.3)}$$

The derivative of the above, with respect to $\beta_i(j)$, is

$$\frac{\partial\ \mathrm{SESE}(j)}{\partial\ \beta_i(j)} = \sum_{i=1}^{n(j)} \frac{\partial\ E\left\{\left(s^{\langle i,j\rangle}\right)^2 \mid \theta'(j)\right\}}{\partial\ \beta_i(j)} - 2\sum_{i=1}^{n(j)} \frac{\partial\ \mu^{\langle i,j\rangle}\ E\left\{s^{\langle i,j\rangle} \mid \theta'(j)\right\}}{\partial\ \beta_i(j)} + \sum_{i=1}^{n(j)} \frac{\partial\ \left(\mu^{\langle i,j\rangle}\right)^2}{\partial\ \beta_i(j)}. \quad \text{(A.4)}$$

In the above, the expectation of $s^{\langle i,j\rangle}$ depends on parameters $\theta'(j)$, not on $\theta(j)$. Therefore it is also independent of $\beta_0(j)$, $\beta_1(j)$, and $\beta_2(j)$, and the first sum in Equation (A.4) equals zero.

To see the similarity with the complete data case, we replace $s^{\langle i,j\rangle}$ with $E\left\{s^{\langle i,j\rangle} \mid \theta'(j)\right\}$, in $\mathrm{SSE}(j)$, given by Equation (4.9), and derivate with respect to $\beta_i(j)$, yielding

$$\sum_{i=1}^{n(j)} \frac{\partial\ \left(E\left\{s^{\langle i,j\rangle} \mid \theta'(j)\right\} - \mu^{\langle i,j\rangle}\right)^2}{\partial\ \beta_i(j)} = \sum_{i=1}^{n(j)} \frac{\partial\ \left(E\left\{\left(s^{\langle i,j\rangle}\right) \mid \theta'(j)\right\}\right)^2}{\partial\ \beta_i(j)} - 2\sum_{i=1}^{n(j)} \frac{\partial\ \mu^{\langle i,j\rangle}\ E\left\{s^{\langle i,j\rangle} \mid \theta'(j)\right\}}{\partial\ \beta_i(j)} + \sum_{i=1}^{n(j)} \frac{\partial\ \left(\mu^{\langle i,j\rangle}\right)^2}{\partial\ \beta_i(j)}. \quad \text{(A.5)}$$

Like in Equation (A.4), the first sum on the right-hand side of Equation (A.5) equals zero. The second and the third sum are identical to the second and third sum of Equation (A.4). Thus, the partial derivatives (A.4) and (A.5) are always equal. The roots of the derivatives are the MLEs of the β-parameters. Consequently, MLEs of the β-parameters in the incomplete data case can be obtained by replacing the unknown variables $s^{\langle i,j\rangle}$ with the their expectations $E\left\{s^{\langle i,j\rangle} \mid \theta'(j)\right\}$, and using the same formula as in the complete data case, given by Equation (4.10).

□

Proof of Equation (4.23)**.** The distribution of $s^{\langle i,j\rangle}$ is assumed to be Gaussian with mean $\mu'^{\langle i,j\rangle}$ and variance $\sigma'(j)^2$, with the additional constraint $o^{\langle i,j\rangle} - \frac{\epsilon}{2} \leq s^{\langle i,j\rangle} \leq o^{\langle i,j\rangle} + \frac{\epsilon}{2}$. Let variable X be defined as

$$X \stackrel{\mathrm{df.}}{=} \frac{s^{\langle i,j\rangle} - \mu'^{\langle i,j\rangle}}{\sigma'(j)}. \quad \text{(A.6)}$$

The distribution of X is Gaussian with zero mean and unity variance. The constraint becomes $a^{\langle i,j \rangle} \leq X \leq b^{\langle i,j \rangle}$, where

$$a^{\langle i,j \rangle} = \frac{o^{\langle i,j \rangle} - \frac{\epsilon}{2} - \mu'^{\langle i,j \rangle}}{\sigma'(j)}, \qquad b^{\langle i,j \rangle} = \frac{o^{\langle i,j \rangle} + \frac{\epsilon}{2} - \mu'^{\langle i,j \rangle}}{\sigma'(j)}. \tag{A.7}$$

The expectation of X is[1]

$$E\{X\} = \frac{1}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j \rangle}) - \Phi(a^{\langle i,j \rangle}))} \int_{a^{\langle i,j \rangle}}^{b^{\langle i,j \rangle}} x \, \exp(-\frac{1}{2}x^2)\, dx, \tag{A.8}$$

where Φ denotes the cumulative distribution function of a Gaussian distribution with zero mean and unity variance. Let functions f and g be defined as

$$f(x) = -\frac{1}{2}x^2, \qquad g(x) = \exp(x). \tag{A.9}$$

Let h be their composite mapping

$$h(x) = (g \circ f)(x) = \exp(-\frac{1}{2}x^2). \tag{A.10}$$

The derivatives of f and g are

$$f'(x) = -x, \qquad g'(x) = \exp(x). \tag{A.11}$$

From Equations (A.9)–(A.11) it follows that the derivative of the composite mapping is[2]

$$h'(x) = \exp(-\frac{1}{2}x^2)\,(-x), \tag{A.12}$$

and thus[3]

$$\int -h'(x)\, dx = \int x \, \exp(-\frac{1}{2}x^2)\, dx = -\exp(-\frac{1}{2}x^2). \tag{A.13}$$

By applying Equation (A.13) to Equation (A.8), we get

$$E\{X\} = \frac{\exp(-\frac{1}{2}(a^{\langle i,j \rangle})^2) - \exp(-\frac{1}{2}(b^{\langle i,j \rangle})^2)}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j \rangle}) - \Phi(a^{\langle i,j \rangle}))}. \tag{A.14}$$

[1] For brevity, we denote the expectation of X by $E\{X\}$, instead of the full notation $E\{\, X \mid X \sim \mathcal{N}(0,1),\ a^{\langle i,j \rangle} \leq X \leq b^{\langle i,j \rangle} \,\}$.

[2] The derivative of a composite mapping is given by $(g \circ f)'(x) = g'(f(x))\, f'(x)$.

[3] We thank Tomi Silander for the above proof of Equation (A.13).

From Equation (A.14), and the definition of X in Equation (A.6), it follows that the expectation of $s^{\langle i,j\rangle}$ is

$$\begin{aligned} E\{s^{\langle i,j\rangle} \mid \theta'(j)\} &= E\{X\}\,\sigma'(j) + {\mu'}^{\langle i,j\rangle} \\ &= \frac{\left(\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - \exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)\,\sigma'(j)}{\sqrt{2\pi}\,\left(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle})\right)} + {\mu'}^{\langle i,j\rangle}. \end{aligned} \tag{A.15}$$

□

Proof of Equation (4.27)**.** The only difference between a truncated and a binned observation is that in the former there is no lower limit for $s^{\langle i,j\rangle}$. Therefore, for a truncated observation, the expectation of $s^{\langle i,j\rangle}$ can be obtained by letting the lower limit, $o^{\langle i,j\rangle} - \frac{\epsilon}{2}$, approach minus infinity. In the limit, both $\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2)$ and $\Phi(a^{\langle i,j\rangle})$ in Equation (A.15) become zero, and the expectation becomes

$$E\{s^{\langle i,j\rangle} \mid \theta'(j)\} = -\frac{\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\,\sigma'(j)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + {\mu'}^{\langle i,j\rangle}, \tag{A.16}$$

where b is given by Equation (A.7). □

Proof of Equation (4.28)**.** The value of $\sigma(j)$ maximizing $Q_j(\theta(j), \theta'(j))$ can be obtained by setting the partial derivative of the latter with respect to $\sigma(j)$ to zero. The first term of $Q_j(\theta(j), \theta'(j))$, given by Equation (4.19), does not contain $\sigma(j)$. Thus, the contribution of the first term to the partial derivative is zero, and the derivative becomes

$$\begin{aligned} \frac{\partial\, Q_j(\theta(j), \theta'(j))}{\partial\, \sigma(j)} &= -\frac{\partial\, n(j) \ln(\sigma(j))}{\partial\, \sigma(j)} - \frac{\partial\, \frac{1}{2}\,\sigma(j)^{-2}\,\mathrm{SESE}(j)}{\partial\, \sigma(j)} \\ &= -n(j)\frac{\partial\, \ln(\sigma(j))}{\partial\, \sigma(j)} - \frac{\mathrm{SESE}(j)}{2}\frac{\partial\, \sigma(j)^{-2}}{\partial\, \sigma(j)}. \end{aligned} \tag{A.17}$$

The two derivatives on the second row have analytical solutions, and we get

$$\frac{\partial\, Q_j(\theta(j), \theta'(j))}{\partial\, \sigma(j)} = -\frac{n(j)}{\sigma(j)} + \frac{\mathrm{SESE}(j)}{\sigma(j)^3}. \tag{A.18}$$

Let $\widehat{\sigma(j)}$ denote the value of $\sigma(j)$ for which the value of the partial derivative is zero. By solving $\widehat{\sigma(j)}$, we get

$$-\frac{n(j)}{\widehat{\sigma(j)}}+\frac{\mathrm{SESE}(j)}{\widehat{\sigma(j)}^{3}}=0 \quad \Leftrightarrow \quad \widehat{\sigma(j)}=\sqrt{\frac{\mathrm{SESE}(j)}{n(j)}}. \tag{A.19}$$

In order to show that $\widehat{\sigma(j)}$ is indeed the maximizer, not minimizer, of $Q_j(\theta(j),\theta'(j))$, we take its the second derivative, which is the derivative of Equation (A.18):

$$-\frac{\partial\, n(j)\;\sigma(j)^{-1}}{\partial\;\sigma(j)}+\frac{\partial\;\mathrm{SESE}(j)\;\sigma(j)^{-3}}{\partial\;\sigma(j)}=\frac{n(j)}{\sigma(j)^2}-\frac{3\;\mathrm{SESE}(j)}{\sigma(j)^4}. \tag{A.20}$$

Substituting $\widehat{\sigma(j)}$, given by Equation (A.19), in the place of $\sigma(j)$, yields

$$\frac{n(j)}{\widehat{\sigma(j)}^{2}}-\frac{3\;\mathrm{SESE}(j)}{\widehat{\sigma(j)}^{4}}=\frac{n(j)^2}{\mathrm{SESE}(j)}-\frac{3\;n(j)^2}{\mathrm{SESE}(j)}=-\frac{4\;n(j)^2}{\mathrm{SESE}(j)}. \tag{A.21}$$

Because both the nominator and the denominator are squared quantities, and therefore, always non-negative, the result is always non-positive. Thus, the second derivative of $Q_j(\theta(j),\theta'(j))$ is non-positive at $\widehat{\sigma(j)}$, which therefore is the maximizer.

□

Proof of Equation (4.29)**.** The distribution of $s^{\langle i,j\rangle}$ is assumed to be Gaussian with mean $\mu'^{\langle i,j\rangle}$ and variance $\sigma'(j)^2$, and with the additional constraint $o^{\langle i,j\rangle}-\frac{\epsilon}{2}\leq s^{\langle i,j\rangle}\leq o^{\langle i,j\rangle}+\frac{\epsilon}{2}$. Let variable X be defined as

$$X \stackrel{\text{df.}}{=} \frac{s^{\langle i,j\rangle}-\mu'^{\langle i,j\rangle}}{\sigma'(j)}. \tag{A.22}$$

The distribution of X is Gaussian with zero mean and unity variance. The constraint becomes $a^{\langle i,j\rangle}\leq X\leq b^{\langle i,j\rangle}$, where

$$a^{\langle i,j\rangle}=\frac{o^{\langle i,j\rangle}-\frac{\epsilon}{2}-\mu'^{\langle i,j\rangle}}{\sigma'(j)}, \qquad b^{\langle i,j\rangle}=\frac{o^{\langle i,j\rangle}+\frac{\epsilon}{2}-\mu'^{\langle i,j\rangle}}{\sigma'(j)}. \tag{A.23}$$

From the definition of X, it follows that

$$E\{s^{\langle i,j\rangle}\mid\theta'(j)\}=\sigma'(j)E\{X\}+\mu'^{\langle i,j\rangle}, \tag{A.24}$$

and

$$E\{(s^{\langle i,j\rangle})^2 \mid \theta'(j)\} = E\{(X\sigma'(j) + \mu'^{\langle i,j\rangle})^2\} \\ = \sigma'(j)^2 E\{X^2\} + 2\sigma'(j)\mu'^{\langle i,j\rangle} E\{X\} + (\mu'^{\langle i,j\rangle})^2. \tag{A.25}$$

The expectation of X^2 is

$$E\{X^2\} = \frac{1}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} \int_{a^{\langle i,j\rangle}}^{b^{\langle i,j\rangle}} x^2 \exp(-\frac{1}{2}x^2)\, dx, \tag{A.26}$$

where Φ denotes the cumulative distribution function of a Gaussian distribution with zero mean and unity variance. Let functions f and g be defined as

$$f(x) = -x, \qquad g(x) = \exp(-\frac{1}{2}x^2). \tag{A.27}$$

From Equation (A.12) it follows that

$$g'(x) = -x \exp -\frac{1}{2}x^2. \tag{A.28}$$

Let $h(x)$ be the product of $f(x)$ and $g'(x)$

$$h(x) = f(x)\, g'(x) = -x\, (-x) \exp(-\frac{1}{2}x^2) = x^2 \exp(-\frac{1}{2}x^2). \tag{A.29}$$

From Equations (A.27) and (A.28) it follows that the integral of the product is given by[4]

$$\int h(x)\, dx = -x \exp(-\frac{1}{2}x^2) + \int \exp(-\frac{1}{2}x^2)\, dx. \tag{A.30}$$

The integrand in the latter term is the density function of a Gaussian distribution with zero mean and unity variance except that it lacks the constant $(2\pi)^{-\frac{1}{2}}$. Therefore the integral can be replaced by $\sqrt{2\pi}\,\Phi(x)$, and the equation becomes[5]

$$\int h(x)\, dx = \int x^2 \exp(-\frac{1}{2}x^2)\, dx = -x \exp(\frac{1}{2}x^2) + \sqrt{2\pi}\,\Phi(x). \tag{A.31}$$

[4] $\int fg'\, dx = fg - \int gf'\, dx$

[5] We thank Tomi Silander for the proof of Equation (A.31) presented above.

From Equations (A.26) and (A.31) it follows that

$$\begin{aligned} E\{X^2\} &= \frac{-b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2) + \sqrt{2\pi}\;\Phi(b^{\langle i,j\rangle})}{\sqrt{2\pi}(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} \\ &\quad - \frac{-a^{\langle i,j\rangle}\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) + \sqrt{2\pi}\;\Phi(a^{\langle i,j\rangle})}{\sqrt{2\pi}(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} \\ &= \frac{a^{\langle i,j\rangle}\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)}{\sqrt{2\pi}(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} + 1. \end{aligned} \tag{A.32}$$

By using Equations (A.24) and (A.25), we obtain a closed form solution for the expectation of $(s^{\langle i,j\rangle} - \mu^{\langle i,j\rangle})^2$ as follows[6]

$$\begin{aligned} &E\left\{\left(s^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)^2 \;\middle|\; \theta'(j)\right\} \\ &\quad = E\{(s^{\langle i,j\rangle})^2\} - 2\mu^{\langle i,j\rangle}E\{s^{\langle i,j\rangle}\} + (\mu^{\langle i,j\rangle})^2 \\ &\quad = \sigma'(j)^2E\{X^2\} + 2\,\sigma'(j)\,\mu'^{\langle i,j\rangle}E\{X\} + (\mu'^{\langle i,j\rangle})^2 \\ &\qquad\quad - 2\,\sigma'(j)\,\mu^{\langle i,j\rangle}E\{X\} + 2\,\mu^{\langle i,j\rangle}\,\mu'^{\langle i,j\rangle} + (\mu^{\langle i,j\rangle})^2 \\ &\quad = \sigma'(j)^2E\{X^2\} + 2\,\sigma'(j)\left(\mu'^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)E\{X\} + \left(\mu'^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)^2. \end{aligned} \tag{A.33}$$

Plugging in $E\{X^2\}$, given by Equation (A.32), and $E\{X\}$, given by Equation (A.14), yields

$$\begin{aligned} &E\left\{\left(s^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)^2 \;\middle|\; \theta'(j)\right\} \\ &\quad = \frac{\sigma'(j)^2\left(a^{\langle i,j\rangle}\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} + \sigma'(j)^2 \\ &\quad + \frac{2\,\sigma'(j)\left(\mu'^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)\left(\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2) - \exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)\right)}{\sqrt{2\pi}\,(\Phi(b^{\langle i,j\rangle}) - \Phi(a^{\langle i,j\rangle}))} \\ &\quad + \left(\mu'^{\langle i,j\rangle} - \mu^{\langle i,j\rangle}\right)^2. \end{aligned} \tag{A.34}$$

□

Proof of Equation (4.30)**.** The only difference between a truncated and a binned observation is that in the former there is no lower limit for $s^{\langle i,j\rangle}$. Therefore, for a truncated observation, the expectation of $s^{\langle i,j\rangle}$ can be obtained by letting the lower limit, $o^{\langle i,j\rangle} - \frac{\epsilon}{2}$, approach minus infinity. In the

[6]For brevity, we use here the short-hand notations $E\{(s^{\langle i,j\rangle})\}$ and $E\{(s^{\langle i,j\rangle})^2\}$, instead of their respective full versions $E\{(s^{\langle i,j\rangle}) \mid \theta'(j)\}$ and $E\{(s^{\langle i,j\rangle})^2 \mid \theta'(j)\}$.

limit, all the terms in Equation (A.34) that are related to $a^{\langle i,j\rangle}$, namely $a^{\langle i,j\rangle}\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2)$, and $\exp(-\frac{1}{2}(a^{\langle i,j\rangle})^2)$, and $\Phi(a^{\langle i,j\rangle})$, become zero, and the expectation becomes

$$\begin{aligned} & E\left\{\left(s^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2 \;\middle|\; \theta'(j)\right\} \\ & \quad = \frac{\sigma'(j)^2\, b^{\langle i,j\rangle}\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + \sigma'(j)^2 \\ & \quad - \frac{2\,\sigma'(j)\left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)\,\exp(-\frac{1}{2}(b^{\langle i,j\rangle})^2)}{\sqrt{2\pi}\,\Phi(b^{\langle i,j\rangle})} + \left(\mu'^{\langle i,j\rangle}-\mu^{\langle i,j\rangle}\right)^2. \quad \text{(A.35)} \end{aligned}$$

where $b^{\langle i,j\rangle}$ is given by Equation (A.23). □