TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

TIINA PARKKINEN

ESTIMATING SOIL FREEZING FROM PASSIVE L-BAND MICROWAVE MEASUREMENTS

Master of Science Thesis

Examiner: D.Sc(Tech) Simo Ali-Löytty
Examiner and subject approved in the
Department Council on 4th June 2014

ABSTRACT

PARKKINEN, TIINA: Estimating soil freezing from passive L-band microwave measurements
Tampere University of Technology
Master of Science Thesis, 71 pages, 8 Appendix pages
November 2014
Master's Degree Programme in Science and Engineering
Major: Mathematics
Examiner: D.Sc(Tech) Simo Ali-Löytty
Keywords: Canny edge detector, frost depth, Kalman filter, multiple linear regression, SMOS satellite

This master of science thesis presents two approaches to estimate soil freezing from passive L-band microwave measurements. The first is the Canny edge detector algorithm which is used to gain estimates on the dates of seasonal changes occurring in the brightness temperature data measured by the ELBARA-II radiometer and the SMOS satellite. Secondly, regression analysis is applied to the ELBARA-II brightness temperature data and frost tube observations to estimate the frost and thaw depths.

The process of soil freezing and the effects of it are shortly introduced. More focus is on the developed and tested observation methods and instruments used at the ground level and by remote sensing methods. From the remote sensing methods the passive L-band ELBARA-II radiometer and the SMOS satellite's MIRAS instrument are discussed in more detail. The Canny edge detector algorithm and the criteria used when deriving it are described in detail. The focus in the multiple linear regression analysis is on the testing and validation of the results obtained. The Kalman filter algorithm used to filter diurnal variations from the brightness temperature data is introduced shortly.

The Canny edge detector algorithm proved out to have a good performance when applied to the local ELBARA-II brightness temperature data. The accuracy was not as good for the SMOS brightness temperature data measured over Sodankylä, since the SMOS data has both lower temporal sampling and larger spatial resolution. Nevertheless, the test performed for the SMOS data measured over Finland in the fall of 2010 showed promising results and expected behavior. All the three approaches chosen for the regression analysis showed similar results, the most accurate being the model where Kalman filtered data was used, while the simplified linearized model and the model fitted using moving average filtered data followed.

TIIVISTELMÄ

PARKKINEN, TIINA : Maaperän routaantumisen estimointi passiivisista L-kaistan mikro-aalto mittauksista
Tampereen teknillinen yliopisto
Diplomityö, 71 sivua, 8 liitesivua
Marraskuu 2014
Teknis-luonnontieteellinen koulutusohjelma
Pääaine: Matematiikka
Tarkastaja: TkT Simo Ali-Löytty
Avainsanat: Canny reunantunnistusalgoritmi, Kalmanin suodatin, lineaarinen regressio, roudan syvyys, SMOS-satelliitti

Tässä diplomityössä esitetään kaksi lähestymistapaa maaperän routaantumisen estimoimiseksi passiivisista L-kaistan mikroaalto mittauksista. Näistä ensimmäinen on Canny reunantunnistus algoritmi, jonka avulla estimoidaan kausittaisten vaihteluiden ajankohdat ELBARA-II radiometrin ja SMOS satelliitin mittaamista kirkkauslämpötiloista. Toiseksi, regressioanalyysin avulla estimoidaan roudansyvyys ja pinnasta sulaneen maaperän syvyys käyttäen ELBARA-II kirkkauslämpötiloja ja routaputkimittauksia.

Maaperän routaantumisen synty ja vaikutukset esitellään lyhyesti. Erityistä huomiota kiinnitetään maanpinnalla sekä kaukokartoituksessa kehitettyihin ja testattuihin havainnointimenetelmiin ja instrumentteihin. Kaukokartoitusmenetelmistä passiivinen L-kaistan ELBARA-II radiometri ja SMOS satelliitin MIRAS instrumentti käsitellään tarkemmin. Canny reunantunnistus algoritmi ja sen johtamisessa käytetyt kriteerit kuvataan perusteellisesti. Monen muuttujan lineaarisen regression esityksessä keskitytään saatujen tulosten testaukseen ja validointiin. Kirkkauslämpötilojen päivittäisten vaihteluiden suodattamiseen käytetty Kalmanin suo-datin esitellään lyhyesti.

Canny reunantunnistus algoritmi osoittautui toimivan hyvin ELBARA-II kirkkauslämpötiloille. Ajallisesti ja alueellisesti harvemmalle SMOS kirkkauslämpötiladatalle Canny reunantunnistus algoritmi ei antanut yhtä tarkkoja tuloksia Sodankylän alueen datalle, mutta koko Suomen yllä syksyllä 2010 mitatuille kirkkauslämpötiloille suoritettu testi käyttäytyi odotetunlaisesti ja antoi lupaavia tuloksia. Kaikki kolme regressioanalyysiin valittua lähestysmistapaa antoivat samankaltaisia tuloksia. Näistä tarkimmat tulokset antoi malli, jossa data suodatettiin Kalmanin suodattimella, seuraavina tulivat yksinkertaistettu linearisoitu malli ja malli, jossa data suodatettiin liukuvalla keskiarvolla.

PREFACE

This master of science thesis has been done at the Finnish Meteorological Institute's Arctic Research Centre in Sodankylä, Finland.

I want to thank my instructor Simo Ali-Löytty from TUT for the guidance and advice given throughout the thesis writing process. I also want to thank my other instructor, Kimmo Rautiainen from FMI, for introducing me to the subject and for his enthusiastic and supportive attitude.

My colleagues Timo Ryyppö and Tomi Karppinen have given precious opinions on the content and styling of tables and figures. My other colleagues in FMI-ARC I want to thank for the technical support and for the joyful conversations, especially in the meetings of KYRPS.

Lastly, I want to thank my friends and my parents for all the support and encouragement given throughout my studies. Especially I want to thank my darling husband Antti for being there for me.

In Sodankylä on the 11th of November 2014

Tiina Parkkinen

CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations

DF	Degrees of freedom
ECMWF	European Centre for Medium-Range Weather Forecasts
ELBARA-II	ETH L-band Radiometer for soil moisture research
ESA	European Space Agency
ESTAR	Electronically Steered Thinned Aperture Radiometer
ETH	Swiss Federal Institute of Technology in Zürich, in German: Eidgenössische Technische Hochschule (ETH) Zürich
FMI	Finnish Meteorological Institute
FMI-ARC	Arctic Research Centre of Finnish Meteorological Institute, located in Sodankylä, Finland
HUT-2D	Helsinki University of Technology 2D radiometer
L	Localization
L-band	Radio spectrum range from 1-2 GHz, corresponds to wavelength of 15-30 centimetres
MIRAS	Microwave Imaging Radiometer using Aperture Synthesis
MSE	Mean square error
MSR	Mean square of regression
MST	Total mean square
NASA	National Aeronautics and Space Administration
RFI	Radio Frequency Interference
RMSE	Root mean square error
SMAP	Soil Moisture Active Passive
SMOS	Soil Moisture and Ocean Salinity
SNR	Signal-to-noise ratio
VLA	Very Large Array radio-astronomy observatory
VLBI	Very-Long-Baseline Interferometry
VWC	Volumetric Water Content

Symbols

∞	Infinity
$\approx$	Approximation
$\|\cdot\|$	Euclidean norm
$\mathbf{0}$	Zero vector
$\mathbf{1}$	A vector with all the entries one
$A \setminus B$	The set of elements in A but not in B
B_f^p	Spectral radiance of an object at frequency f and at polarization p
$\mathrm{C}_x(\tau)$	Autocovariance of x
δ_e	Emission depth
$\delta(\tau)$	The Dirac delta function
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$	Differentiation of function f with respect to x
ϵ	Dielectric permittivity of soil
$\mathrm{E}(x)$	Expectation value of x
exp	Exponential function
$f'(x)$	Derivative of function f
$f''(x)$	Second derivative of function f
$(f * g)(x)$	Convolution of functions f and g
$F(m, n)$	F-distribution with m and n degrees of freedom
FD	Frost depth
FF	Frost factor
$\int_a^b f(x)\mathrm{d}x$	Definite integral of function f over the interval $[a, b]$
$\int f(x)\mathrm{d}x$	Definite integral of function f over the unbounded interval $(-\infty, \infty)$
I	Identity matrix
$\Im(z)$	Imaginary part of a complex number z
log	The natural logarithm function
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
O	Landau's symbol for describing the limiting behavior of a function
$P(x)$	Transmitted power of radiation in a medium at depth x
$\mathbb{P}(x)$	The probability of event x
p	Polarization, notation $p = \mathrm{H}$ for horizontal and $p = \mathrm{V}$ for vertical polarization
R^p	Surface reflectivity at polarization p
$\mathbb{R}$	The set of real numbers

$\mathbb{R}^{n\times m}$	The set of real $(n \times m)$ -dimensional matrices
$\mathrm{R}_x(\tau)$	Autocorrelation of x
SS	Sum of squares
T_B^p	Brightness temperature at polarization p
$t(m)$	t-distribution with m degrees of freedom
$t_{\frac{\alpha}{2}}$	A quantile of the standard normal distribution so that for a random variable t $\mathbb{P}(t \geq t_{\frac{\alpha}{2}}) = \alpha/2$
T_S	Surface temperature
T_{sky}	Downwelling sky radiance
$u_{-1}(x)$	Unit step function
x, y	A random variable or a vector
$x \sim X$	The random variable x follows a distribution X
$\hat{x}$	An estimate of x
$\bar{x}$	The mean of x
X	A matrix
X^{-1}	The inverse of matrix X
X^T	The transpose of matrix X
$\mathrm{V}(x)$	Variance of x

1. INTRODUCTION

The landmass affected by the soil freezing, either seasonal or perennial, is significant, being more than half of the total landmass of the Earth. Frozen soil alters the soil's thermal and hydrological properties, affects the ecosystem diversity and productivity, the global carbon cycle and the evapotranspiration of different atmospheric gases. The water, energy and carbon cycles, which are all greatly affected by the soil freezing, are all important factors in numerical models of the atmosphere used in climate and weather prediction. Therefore a good knowledge of soil's thermal state could improve the performance of these atmospheric models as well as hydrological and climatological forecasts. Also drought monitoring and agricultural predictions would benefit from this information.

The freezing and thawing of the soil can be monitored at ground level with many different methods including soil sampling and by conducting frost tube measurements. These give a very local, but quite accurate estimate of the thermal state of the soil. Because these methods give very local information and are time consuming, several remote sensing methods have been tested in the past decades. Already since 1978, microwave measurements with radars have been conducted with fine resolution but with low temporal sampling in the order of tens of days. The most recent focus has been on the use of low frequency radiometers working at L-band of 1-2 GHz reserved for radio-astronomy. This frequency band is chosen because the low frequency measurements are not affected much by the clouds, vegetation or atmosphere. Two satellite missions with radiometers measuring at L-band are currently on-going, the already operational Soil Moisture and Ocean Salinity (SMOS) mission by European Space Agency (ESA) and the Soil Moisture Active Passive (SMAP) mission by National Aeronautics and Space Administration (NASA), scheduled for launch in January 2015.

At these low frequencies the emissivity of land is a strong function of soil moisture. This originates from the high dielectricity of water which in turn gives rise to a contrast between dielectricity of wet and dry soils. There is also a contrast between the dielectricity of free water and ice, which makes it possible to detect the thermal state of the soil. The distinction between dry soil and frozen soil is not

that straightforward and, furthermore, the potential coexistence of ice, liquid water and bound water in soil and a possible snow cover have an additional effect to the detected emission.

As measurements made from space are always indirect, the remote sensing instruments are calibrated and validated using in situ ground level measurements that need to be frequent and representative. One such a ground level calibration and validation site has been set up in Sodankylä, Finland, to Finnish Meteorological Institute's Arctic Research Centre (FMI-ARC). One of the official reference instruments for the SMOS satellite, the ELBARA-II radiometer, is also located there. The ELBARA-II is a passive microwave radiometer that measures thermal radiation emitted from the surface of the Earth. The ELBARA-II measurements are accompanied with in situ measurements of the frost depth with frost tubes and other climatological data.

To gain estimates on the start dates of the freezing and thawing periods, the Canny edge detector algorithm can be applied to the brightness temperature data measured by the ELBARA-II radiometer and the SMOS satellite. The Canny edge detector algorithm was originally formulated for identifying features in noisy 2D-images, but it can be applied also for noisy one dimensional data to detect significant local changes in the signal. The performance of the Canny edge detector algorithm will also be tested on a larger scale for the SMOS brightness temperature data measured over Finland during autumn 2010.

In an earlier study done at the FMI-ARC, an algorithm was formulated to detect the thermal state of the soil by estimating the frost depth from the ELBARA-II measurements [38]. In this algorithm, the data was filtered with moving average filter because of the large diurnal variations present in the brightness temperature data. As the moving average filter places equal emphasis on all data points inside the time span chosen, the values in the past will have the same influence as a more current measurement. One way to overcome this, is to filter the data with the Kalman filter, which places more emphasis on the most recent measurements. The model formulated in [38] is nonlinear and was optimized using the freezing period data. The freezing and thawing periods can also be modeled separately. This approach, where two simpler, linear models are formulated, can be obtained by using multiple linear regression analysis. The significance of the model and the assumptions made when formulating it can be analyzed by hypothesis testing and residual analysis.

The contents of this master of science thesis are organized as follows: Chapter 2 focuses on describing the term thermal state of the soil and ways to detect it both at ground level and by using remote sensing methods. The ELBARA-II radiometer and the SMOS satellite mission are also discussed in this chapter. In Chapter 3, the Canny edge detection algorithm and the criteria used for the edge detector performance are defined. The Kalman filter algorithm is shortly introduced and the state and measurement models formulated. The multiple linear regression analysis is also described in this chapter, the focus being on the testing and validation of the results obtained. The results of applying the Canny edge detector algorithm and the regression analysis are given in Chapter 4. Lastly the results are summarized and thoughts on further research are given in Chapter 5. Some figures of both the Canny edge detection results and regression analysis are shown in Appendix A and Appendix B.

2. RETRIEVING THE THERMAL STATE OF THE SOIL

Around 70 percent of the Earth's surface is covered by water. From the remaining landmass more than half is affected by soil freezing, either seasonal or perennial. As soil freezes, it affects the water, energy and carbon cycles, which are all important factors in numerical models of the atmosphere used in climate and weather prediction. More practical impacts of freezing and thawing of soils include damage caused to construction and roads due to frost heaving, whereas in early spring the timing of cultivation and seeding is dependent on the thaw depth. Therefore a good knowledge of soil's thermal state could improve the performance of atmospheric models as well as hydrological and climatological forecasts, drought monitoring and agricultural predictions.

Since the soil's freezing and thawing events play such an important role, this chapter is focused on explaining the term thermal state of the soil, describing ways to detect the current thermal state both at ground level and by remote sensing methods and justifying the need for such a information. Also the related ELBARA-II instrument, located in FMI-ARC, and the SMOS satellite mission are explained in more detail in this chapter.

2.1 The thermal state of the soil

As mentioned above the majority of the Earth's surface is covered by water. This mass of water is constantly moving through the surface-atmosphere interface by several different physical processes, including condensation, precipitation, transpiration and evaporation. At the surface water can be stored in liquid or solid form in vegetation, oceans, lakes, rivers, snow, ice and soil. This mass transfer of water, described by the global water cycle, has a great impact on the current and future weather conditions and local surface water content. Excess of water, due to snow melting or heavy rainfalls, causes floods and landslides, on the other hand shortage causes droughts that have a direct impact on plant growth and crop productivity

and thus hunger. In addition, snowmelt runoff has effect on ocean salinity and sea ice conditions, latter being a major risk to ship traffic and oil drilling.

Besides mass water also transfers energy as heat especially in running water due to its high specific heat capacity. Evaporation of water has a special significance since it requires energy which in turn induces a decrease in temperature and therefore it affects the global energy budget at surface-atmosphere interface [19, p. 677]. A factor that controls the partitioning of energy available in the regions where the evaporation regime is water-limited is soil moisture [9, p. 705].

In winter time more than half of the Earth's landmass is effected by soil freezing. Frozen soil alters the soil's thermal and hydrological properties by delaying the cooling of the land surface in winter when the soil freezes and also delaying the summer warming when the soil thaws. Frozen soil affects also the ecosystem diversity and productivity, as seasonal freezing and thawing of soil causes changes in the structure of the soil and regulate the availability of nutrients needed by plants and biota. Also the carbon cycle and the evapotranspiration of different atmospheric gases are affected by soil freezing. The gas exchange between the atmosphere and land surface is increased after thawing of the soil, being minimal when the soil is frozen. The situation is the same for perennial areas which are thawing as the climate is gradually warming. [32, p. 937]

The effects of the thermal state of the soil, which refers to the information whether the soil is frozen or not, should now be apparent, but how is frozen soil actually defined? As pure water freezes at the temperature of 0 °C, this could be thought of as a threshold for soil starting to freeze as well. However, soil is a mixture of minerals, organic matter and pores of gases and liquids, which all have an influence to the soil freezing process. The exact composition and texture of the soil, the soil type, can vary widely even on a few meter scale. The present soil type affects the amount of water absorbed by the soil before all the available pores in the soil are filled with water, i.e. it saturates. Water has high specific heat capacity, which causes the soils with higher water content to freeze later than soils with lower water content. All in all, the water content in the vadose zone, defined as the soil between the land surface and the water table, can vary considerably. Furthermore, the varying salinity of the water in the soil has an effect of lowering the actual freezing point of the soil. That is why frozen soil is usually defined as soil which contains ice. The process when frozen soil melts is referred to as thaw. It is possible to measure the soil frost and thaw depth, defined as the distance from the ground level to the freezing front of soil, and from the ground level to the thawing front of soil, respectively.

To gain the soil moisture and soil's thermal state estimates with adequate temporal and spatial resolution, numerous studies and missions have been conducted and several are currently ongoing, for example [9, 18, 19]. One of these missions, the SMOS satellite mission, and a few of the ground level measurement methods are discussed in more detail next.

2.2 Measurements at ground level

At ground level many of the methods commonly used include hard work in the form of digging a hole into the ground either to collect a soil sample for direct observations or to be able to install sensors into the ground. With direct observations frozen soil is distincted from unfrozen according to the soil's hardness or by the existence or absence of visible ice. This method is highly destructive, thus time series of observations cannot be made from the exactly same location or nearby where the soil surface has been disturbed by previous soil sampling. The procedure can also be time consuming and rather challenging, especially if the frost level is deep or the ice content of the soil high. Nevertheless, the observations made can be very descriptive and accurate on a local scale.

Another option is to install automatic temperature sensors into the ground in several depths and estimate the frost and thaw depth from the measured soil temperature. This method has the advantage of giving better temporal and spatial resolution as automatic sensors can be installed to many locations, but on the other hand several sensors have to be installed on top of each other, say with 10 centimeter spacing, to gain the temperature profile measurements needed to estimate the frost depth. Since the sensors are installed in discrete depths the soil temperature between two measurement depths needs to be estimated and is subject to errors that depend on the sensor spacing and interpolation and extrapolation method used [40]. Also the correspondence between the soil temperature with the freezing temperature of that specific soil type needs to be determined since the soil may not be frozen even though the temperature would be below 0 °C, as discussed earlier.

A commonly used instrument for monitoring frost and thaw depth is called a frost tube, shown in Figure 2.1. A frost tube, which is installed vertically into the ground, has an inner acrylic tube filled with methylene blue solution and an outer guide of polyvinyl chloride (PVC) pipe. Since methylene blue solution changes color from blue to colorless when it freezes, the frost and thaw depths can be measured by pulling the inner tube upwards from the outer guide and recording the length of the colorless section. An example of the measured frost and thaw depths in an open forest site are shown in Figure 2.2.

Figure 2.1 *A frost tube with the inner tube inside the outer guide without the protective cap.*

According to [16, p. 116], the frost depths measured using a frost tube are accurate and the method itself is both convenient and inexpensive. A considerable caveat is the overestimation of thaw depth, resulting presumably from the heat conduction along the frost tube and the early melting of snow cover near the frost tube. [16, p. 111-113]

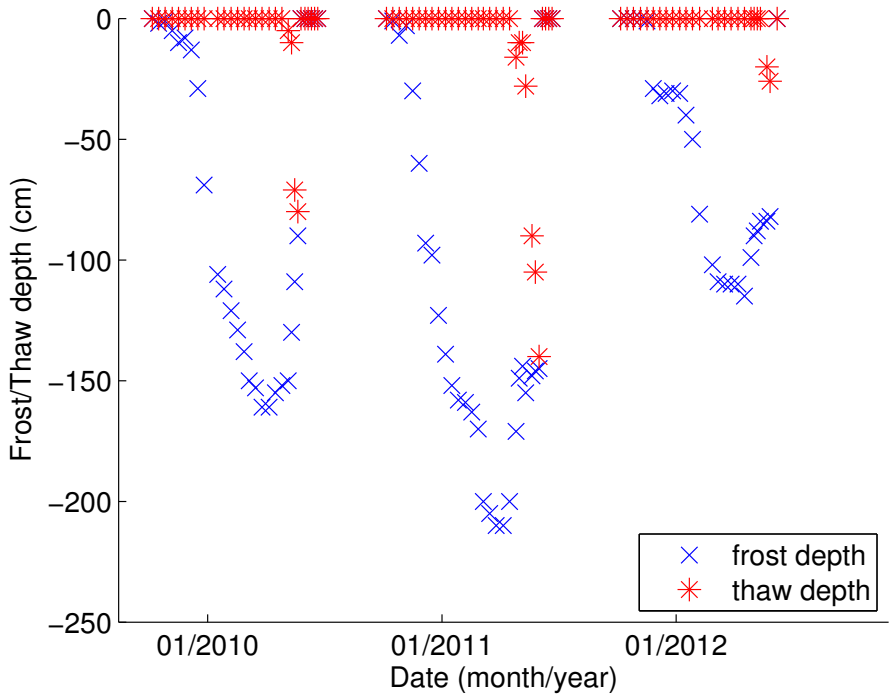

Figure 2.2 *Frost and thaw depths measured in an open forest site. The measurement interval was 10 days.*

If direct observations, frost tube measurements or soil temperature data are not available, frost depth can be estimated using air temperature data as soil temperature correlates relatively well with air temperature. Air temperature is measured and also estimated globally. The globally estimated air temperature data is available e.g. from the European Centre for Medium-Range Weather Forecasts (ECMWF) database [10]. In the literature there are numerous formulas for calculating freezing and thawing indices used when predicting and mapping permafrost distributions and to predict the frost depth [11, p. 47-48]. Air temperature data can also be used together with additional snow depth data to estimate frost depth up to a few centimetres accuracy [31].

Still one approach would be utilizing automatic soil moisture sensors installed into the ground to detect the freezing and thawing events of the soil. One of the approaches used when measuring soil moisture is to measure the dielectric permittivity ϵ of the soil, which can be related to the water content of the soil. The dielectric permittivity of a dry soil differs from that of a wet soil due to the high permittivity of water. There has been many studies about the relationship between dielectric permittivity and the water content of the soil. Therefore there are numerous conversion equations in the literature, see e.g. [47, 4], for different soil types which can be used when transferring measured dielectric permittivity to the water content of that specific soil type. The caveats of this method are similar to those of the automatic temperature sensors'.

Because many of the ground level measurements of frost and thaw depth lack a level of representativeness, there has been and are still ongoing many satellite missions to get global estimates of frost and thaw depth from the space with remote sensing methods. An overview to these and a more thorough description of the ELBARA-II instrument and the SMOS satellite mission are given next.

2.3 Remote sensing methods

Ground level measurements of frost depth can be performed inexpensively and conveniently but a need of global estimations of the thermal state of the soil and other quantities such as soil moisture with adequate temporal sampling has led to a vast testing of remote sensing approaches during the past decades. These include various microwave instruments operating in different frequency bands, instruments using latent heat effects and a newer approach relying on measurements of the gravity field.

The first measurements for surface soil moisture were short-wave measurements utilizing the fact that wet soils are darker in color than drier soils. However, due to potential cloud and vegetation cover and atmospheric effects this was not a very sensitive method. In another approach the premise was the latent heat effects of the soil. The higher thermal inertia of wet soil was determined by, amongst all, thermal inertia monitoring, monitoring the rate of heat in the morning and measuring the amplitude of surface temperature. Suffering from the same effects than the short-wave measurements and the need of additional information, such as wind conditions, proved these approaches to be disappointing. The gravity field measurements are relying on the fact that gravity can be straightly linked to mass and on short time scales changes in mass are linked to mass transfer caused by moving water. The mass transfer of water includes the water in lakes, snow and ice, water in vegetation and also ground water and thus gravimetry indicates the changes in the total column of water. The relationship with water storage and the signal output needs to be validated and explained still, and a major problem in this method is the need for a large number of corrections. [19, p. 669]

Already since 1978 microwave measurements with radars have been conducted with fine resolution but with low temporal sampling in the order of tens of days. For scatterometers the spatial resolution is coarser but the sampling is much more frequent with around 4-6 days. The most recent focus has been on the use of low frequency radiometers with more frequent temporal sampling but with a coarser resolution of around 50 km. The low frequency microwave measurements are performed at L-band of 1-2 GHz, more precisely at the narrow frequency band from 1400 to 1427 MHz, which is reserved for radio-astronomy. Optimistically this band should not have any man-made interferences, since it is protected by international radio regulations adopted by the World Radiocommunication Conference of the International Telecommunications Union. Additionally low frequency measurements are not affected much by the clouds, vegetation or atmosphere. [19, p. 670]

At these low frequencies the emissivity of land is a strong function of soil moisture [19, p. 667]. This originates from the high dielectricity of water which in turn gives rise to a contrast between dielectricity of wet and dry soils. There is also a contrast between the dielectricity of free water and ice, which makes it possible to detect the thermal state of the soil. However, the distinction between dry soil and frozen soil is not straightforward as frozen soil affects the detected emission in the same way as dry soil does, i.e. increases the emission detected. Furthermore, there might be coexistence of ice, liquid water and bound water in soil. The possible snow cover has an additional effect on the emission detected, as there is refraction from the snow surface. [38, p. 207]

The radiometer receives the total radiance, or brightness, B_f^p emitted from the surface of the Earth. The radiance B_f^p emitted at horizontal or vertical polarization, $p = \mathrm{H}$ or $p = \mathrm{V}$ respectively, depends on the surface reflectivity R^p which is the integral of the surface scattering coefficient κ_s over all scattering directions, and on the surface temperature T_S. When the radiometer is oriented towards the surface, the radiance received can be expressed by

$$B_f^p = (1 - R^p)T_S + R^p T_{\mathrm{sky}},$$

where the $(1 - R^p)T_S$ part corresponds to Rayleigh-Jeans approximation of the Planck function of thermal radiation which is linear with the absolute temperature in the microwave range, and T_{sky} is downwelling radiation determined by the cosmic background temperature of around 2.7 K enhanced by an atmospheric contribution. From the surface reflection coefficient R^p, the dielectric constant ϵ of the soil can be estimated using Fresnel reflection equations

$$R^{\mathrm{H}} = \left| \frac{\cos\theta - (\epsilon - \sin^2\theta)^{\frac{1}{2}}}{\cos\theta + (\epsilon - \sin^2\theta)^{\frac{1}{2}}} \right|^2 \quad \text{and} \quad R^{\mathrm{V}} = \left| \frac{\epsilon\cos\theta - (\epsilon - \sin^2\theta)^{\frac{1}{2}}}{\epsilon\cos\theta + (\epsilon - \sin^2\theta)^{\frac{1}{2}}} \right|^2,$$

where θ is the viewing angle. From the Fresnel equations it follows, that the contrast of dielectric constants between two media will affect the two linear polarizations differently. For frozen soil the difference of emitted signals between polarizations can be expected to be smaller than for unfrozen soil. For the case of a dry snow-cover on the ground, there is refraction at the air-snow interface, which decreases the viewing angle at the snow-soil interface according to Snell's law. This decrease in the viewing angle is found to cause a change in the detected emissivity at horizontal polarization, while the emissivity at vertical polarization shows less sensitivity to snow-cover at a viewing angle of 50° [42, p. 187-188]. [38, p. 207]

As T_S represents the effective temperature of the soil averaged over the emission depth δ_e of the radiation, the emission depth, defined as the depth beyond which the power of the emitted signal has decreased by e^{-1} of its original value, is relevant. The radiative transfer model is widely used in remote sensing in the modeling of the propagation of microwave signals in natural media. The theory describes the propagation of electromagnetic intensity in a media affected by absorption, emission and scattering. The general model is rather complex but in remote sensing a typically adopted assumption of planar waves propagating in a media homogeneous in the azimuth direction, representing the atmosphere, simplifies the model. In Carte-

sian coordinates the model takes the following differential form for a plane wave propagating in the direction θ_i respect to the z-axis

$$\begin{aligned}\frac{\mathrm{d}B_f(z,\theta_i)}{\mathrm{d}z} =& -\kappa_e(z)\sec\theta_i B_f(z,\theta_i) + \kappa_s(z)\frac{\sec\theta_i}{2}\int_0^\pi \Psi(\theta')B_f(z,\theta')\sin\theta'\mathrm{d}\theta' \\ &+ \kappa_a(z)\sec\theta_i J(z),\end{aligned} \tag{2.1}$$

where B_f is the affecting brightness, κ_e is the extinction coefficient representing total loss of energy in the medium, κ_s is the scattering coefficient, Ψ is the scattering phase function, κ_a is the absorption coefficient and J is the emission source function characterizing the medium. In the spectral band of Δf, the emission source function can be expressed using Rayleigh-Jeans approximation by

$$J = \frac{2k_B}{\lambda^2}T\Delta f,$$

where k_B is the Boltzmann constant and λ is the wavelength of the emitted signal. With this substitution the radiative transfer equation (2.1) can be expressed in terms of brightness temperature T_B as

$$\begin{aligned}\frac{\mathrm{d}T_B(z,\theta_i)}{\mathrm{d}z} =& -\kappa_e(z)\sec\theta_i T_B(z,\theta_i) + \kappa_s(z)\frac{\sec\theta_i}{2}\int_0^\pi \Psi(\theta')T_B(z,\theta')\sin\theta'\mathrm{d}\theta' \\ &+ \kappa_a(z)\sec\theta_i T(z).\end{aligned} \tag{2.2}$$

This is a linear first-order differential equation with respect to z [15, p. 161]. The brightness temperature at a distance H in the media can be solved from (2.2), and can be written as [50, p. 81]

$$\begin{aligned}T_B(H) =& \, T_B(0)e^{-\kappa_e H\sec\theta_i} \\ &+ \sec\theta_i\int_0^H\left[\frac{\kappa_s(z')}{2}\int_0^\pi \Psi(\theta')T_B(z',\theta')\sin\theta'\mathrm{d}\theta' + \kappa_a(z')T(z')\right]e^{-\kappa_e(z')\sec\theta_i}\mathrm{d}z'.\end{aligned}$$

From this equation it can be seen that the radiation, and thus the observed brightness temperature, decays exponentially with respect to the distance traversed in the media. [24, p. 40,43-44]

The observed emission comes from the entire soil profile but much of the energy is absorbed by the upper layer of the soil only. The effective emission depth depends on the water content and physical properties of the soil and the variation is considerable as it can vary from one meter for dry soils to only a few centimetres for wet soils. New algorithms are needed to determine the emission depth since it is still an important issue not yet entirely solved.

To obtain an expression for the theoretical emission depth δ_e, consider a plane wave incident upon a soil surface from the air in the direction of z. A part of the power is scattered back into the air, while the remainder is transmitted into the soil. Marking the transmitted power just beneath the soil surface with $P(0)$, the power at a depth z is can be expressed as [49, p. 847]

$$P(z) = P(0) \exp\left(- \int_0^z \kappa_e(z') \mathrm{d}z' \right), \tag{2.3}$$

where $\kappa_e(z)$ is the extinction coefficient at depth z expressing how the intensity of radiation traversing through a medium is reduced. As was mentioned earlier, the emission depth is defined as the depth beyond which the power of the emitted signal has decreased by e^{-1} of its original value. Thus, using (2.3) we have

$$P(\delta_e) = \frac{1}{e} P(0) \iff \int_0^{\delta_e} \kappa_e(z') \mathrm{d}z' = 1. \tag{2.4}$$

Approximating $\kappa_e(z)$ to be a constant, δ_e can be solved from (2.4) to be

$$\delta_e = \frac{1}{\kappa_e}. \tag{2.5}$$

The power lost from the incident radiation may have been absorbed by the medium, scattered, or both. Since absorption and scattering are linear processes [48, p. 212], the extinction coefficient κ_e can be expressed as the sum of the absorption coefficient κ_a and the scattering coefficient κ_s as

$$\kappa_e = \kappa_a + \kappa_s. \tag{2.6}$$

The absorption coefficient κ_a can be expressed by using the attenuation coefficient α [48, p. 225] as

$$\kappa_a = 2\alpha = 2k_0 |\Im(\sqrt{\epsilon})| = \frac{4\pi |\Im(\sqrt{\epsilon})|}{\lambda_0}, \tag{2.7}$$

where k_0 and λ_0 are the wave number and wavelength in free space respectively. If we ignore scattering in the soil, $\kappa_e \approx \kappa_a$ according to (2.6) and (2.5) can be written using (2.7) as

$$\delta_e = \frac{1}{\kappa_e} \approx \frac{1}{\kappa_a} = \frac{\lambda_0}{4\pi |\Im(\sqrt{\epsilon})|}. \tag{2.8}$$

By using experimental data and (2.8), [37] estimated the dielectric constant for the frozen soil in FMI-ARC site to be $3.6 - 0.9$j, resulting in a theoretical emission depth of 7.3 cm, while in the literature the permittivity values of frozen soil are typically between $4 - 0.5$j and $6 - $j resulting in a theoretical emission depths of 13.7 cm and 8.4 cm [13, p. 32]. [38, p. 207]

Two satellite missions with radiometers measuring at L-band are currently on-going, the already operational Soil Moisture and Ocean Salinity (SMOS) mission by European Space Agency (ESA) and the Soil Moisture Active Passive (SMAP) mission by National Aeronautics and Space Administration (NASA), scheduled for launch in January 2015. SMAP also utilizes a L-band radar onboard to be able to offer measurements of surface soil moisture and freeze and thaw state of the soil with spatial resolutions of 40 and 3 km and a three-day global revisit time [21, p. 735]. The SMOS mission is described in more detail in Section 2.3.2.

Even though the frequency band from 1400 to 1427 MHz is protected, there is still a possibility for the signal to be affected by man-made emissions and radio frequency interference (RFI), transmission and leakage. For this reason for example the SMAP mission has adopted RFI detection and mitigation measures [9, p.707]. [19, p. 667,669-670]

As measurements made from space are always indirect and can be contaminated by man-made emissions, the remote sensing instruments are calibrated and validated using in situ ground level measurements that need to be frequent and representative of various soil types. Since the need for these ground level measurements is significant, one such a ground level calibration and validation site has been set up in Sodankylä, Finland, to FMI-ARC. One of the official reference instruments for SMOS satellite is also located there. This ELBARA-II instrument is described in more detail next.

2.3.1 ELBARA-II instrument

Before launching an satellite and also during the commissioning phase and the operative use, the instrument onboard the satellite and the models and algorithms used in the processing of the measurement data are tested. Also validating the satellite data with in situ data is highly important. For the validation purposes of European Space Agency's (ESA) Soil Moisture and Ocean Salinity (SMOS) mission three identical L-band radiometers were ordered. One of the three reference instruments is located in FMI-ARC in Sodankylä, Finland. This dual polarization ETH (Swiss Federal Institute of Technology in Zürich) L-band radiometer for soil moisture research,

ELBARA-II, is mounted on a 5 meter tall observation tower. An overall image of the instrument is shown in Figure 2.3.

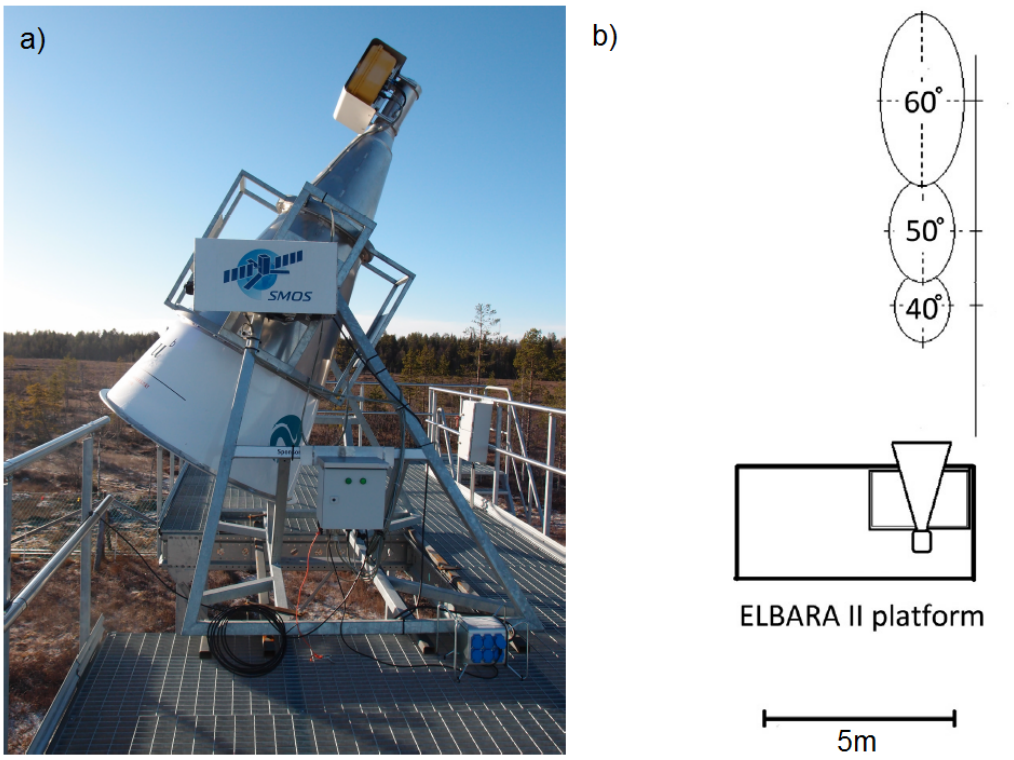

Figure 2.3 *a) The ELBARA-II radiometer standing in a tower at the bog site. b) Schematic figure of the ELBARA-II radiometer and footprint locations at some of the observation angles. Modified from Fig.2. [38, p.207].*

ELBARA-II is a passive microwave radiometer that measures thermal radiation emitted from the surface of the Earth. In FMI-ARC, the ELBARA-II instrument was situated in a forest opening for three consecutive winters, after which it was moved to a nearby bog site in August 2012. The daily measurement procedure consists of an 30 minute's elevation scan on inclination angles from 30° to 70° in 5° steps every four hours, an 120 minutes fixed 50° angle measurement between the scans and a short 10 minute sky measurement conducted every night to verify receiver's stability. During the first year the elevation scans were performed every three hours. One measurement for an individual observation angle takes 150 s, with an recommended integration time of 3 s and resulting in uncertainty of less than 0.1 K [43, p. 606]. A schematic image of the observation angles is shown in Figure 2.3.

The elevation scan data and calibration data received from the radiometer are further processed in FMI-ARC to brightness temperatures T_B on horizontal and vertical polarizations. A three year time series of the calibrated data with a measurement angle of 50 ° is shown in Figure 2.4.

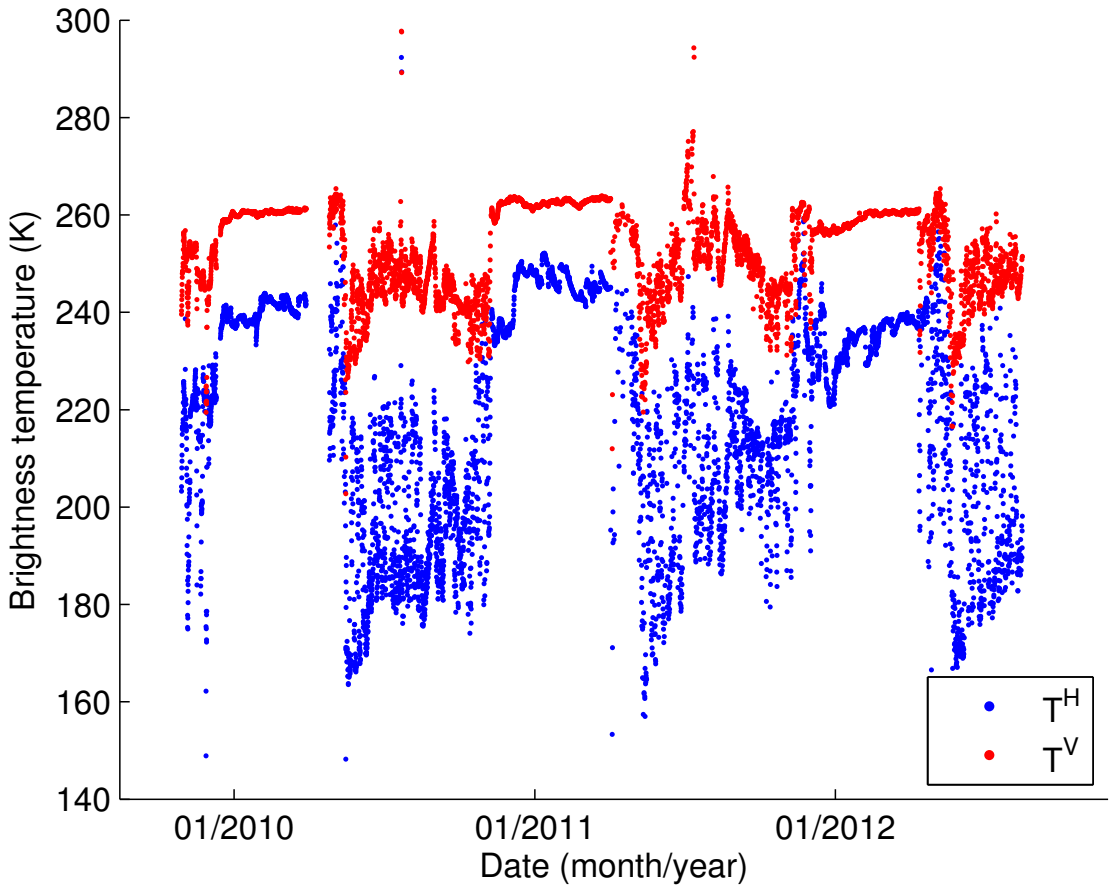

Figure 2.4 *Calibrated data from the ELBARA-II radiometer from 50° measurement angle in horizontal and vertical polarizations. Outliers have not been removed. Majority of the outliers lie out of the axis limits.*

As can be seen from the data, the brightness temperatures for vertical polarization are larger than those of horizontal. During winter periods T_B is higher and has decreased variability at both polarizations than during summer periods. As expected, the polarization difference is larger during summer periods than during the winter. It has been demonstrated that the increase in T_B in autumn is clearly associated with soil freezing [37]. The connection between thawing and the decrease of T_B in the melting period is not that clear, since there is liquid water present in the soil surface and also in the overlying snow. Even though dry snow is relatively transparent at L-band, the local refraction angle inside the snow cover can vary due to melt-refreeze crusts and thus alter the detected emission. [38, p. 210,214]

The calibrated data from ELBARA-II instrument is used in the validation of the SMOS satellite data as mentioned earlier. The SMOS satellite mission is discussed in more detail next.

2.3.2 SMOS satellite

As the need for global estimates on soil moisture and ocean salinity are required for improving meteorological and climate predictions, ESA has selected the SMOS mission as its second Earth Explorer Opportunity Mission [19, p. 666]. The SMOS satellite, launched on November 2009, carries an L-band 2-D interferometric radiometer, which operates in the 1400-1427 MHz protected microwave band. It receives the radiation emitted from the surface of the Earth, which can be related to the water content of the soil over land and sea-surface salinity over the oceans. An image of the satellite is shown in Figure 2.5.

Figure 2.5 *An artistic view of the SMOS satellite, showing the Y-shaped form of the instrument [45].*

The satellite is on a sun-synchronous orbit which combines the altitude and inclination in such a way that the satellite ascends or descends over the equator at the same local mean solar time. For SMOS it is at 6am and 6pm local solar time. The globe is fully imaged twice every three days at an average resolution of 43 kilometres. The requirement for the soil moisture accuracy is 0.04 m^3/m^3 or better and the freeze/thaw transitions should be captured with two-day precision. Any point on the surface of the Earth is imaged from several angles as the satellite travels on its orbit in an altitude of 758 km. To avoid potential biases, the exact repeat time of the orbit is 149 days to ensure that the surface is rarely seen from the exactly same angle. [19, p. 673-674]

The sole payload of the SMOS satellite is the Microwave Imaging Radiometer with Aperture synthesis, abbreviated with MIRAS. It consists of 69 elementary receivers and small antennas regularly spaced in three arms in a Y-shape form. Each of the receivers measure the brightness temperature through a wide-beam patch antenna. There is a phase difference between electromagnetic waves received by different antennas a known distance apart from each other. As can be seen in the Figure 2.6, the receivers form each an interferometric pair, a baseline, with all of the other receivers. This is why the aperture synthesis technique used in the MIRAS instrument is also called the interferometric measurement technique as the image of the target is a synthesis of all possible interferometric patterns of two receivers. The baselines are correlated with one another [24, p. 50] and all the receivers are provided with a correlated noise test signal to ensure the correct correlation and optimum performance [27, p. 595-596]. Due to the interferometric measurement principle, the resulting field of view is a hexagon-like shape of the size of about 1000 km $\times$ 12000 km. Every three seconds such a field of view is obtained with an approximate displacement of 22 km. As a result a point on the Earth's surface is measured a number of times from different angles as the satellite travels on its orbit [20, p. 1730]. The configuration of the instrument provides a full image at either two polarizations at each integration step or a full polarization of the Earth's surface.

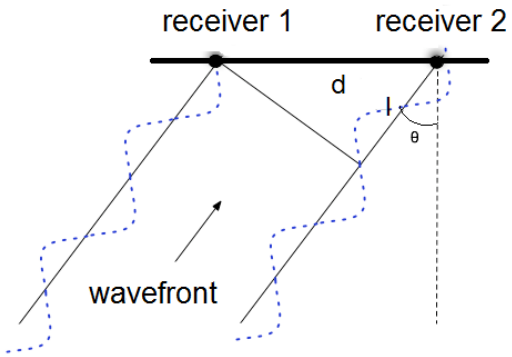

Figure 2.6 *Schematic image of the interferometric measurement principle. The spacing of the receivers is d, the wavefront incidence angle θ and the phase-difference l.*

The aperture synthesis technique is widely used in radio-astronomy in Very-Long-Baseline Interferometry (VLBI) arrays and Very Large Array (VLA) radio-astronomy observatory for investigation of point targets such as the sun, planets, galaxies, quasars and black holes. For airborne remote sensing the aperture synthesis technique was introduced and successfully tested first time in one dimension with Electronically Steered Thinned Aperture Radiometer (ESTAR) instrument in NASA. The first successful aperture synthesis airborne measurements in two dimensions

were conducted in 2005 with Helsinki University of Technology 2D radiometer (HUT-2D), with which some of the technological solutions chosen for the MIRAS concept were verified in practice [36, p. 717]. However, for global Earth remote sensing with wide targets providing different challenges for instrument calibration, the SMOS satellite was the first one using this technique. [29, 30, 36]

The measured raw data from the SMOS datellite is processed in the ground segment into different products. The brightness temperature product is generated with a very complex procedure based on the measurement of cross-correlations of interferometric pairs formed by the receivers. The correlations are then used to form the calibrated complex visibility function which is inverted using inverse Fourier transform to get the brightness temperature of the measured scene [3, p. 1126-1127,1132]. In this process any inaccuracy in the measurement or change in the antenna directivity are directly converted to retrieval errors [25, p. 3752]. The obtained brightness temperatures are averaged over five degree intervals and the standard deviation of the brightness temperatures over that interval is calculated.

The measurement approach requires careful calibration scheme which is relying on on-ground characterization performed before the launch and regular calibration measurements during operative mission consisting of several procedures to calibrate different elements of the instrument [24, p. 125]. The calibration of the instrument is twofold as the noise injection radiometers require classical calibration approach, whereas the interferometer needs novel approaches. During the commissioning phase the optimal rate for the different calibration activities required to remove instrumental errors were determined. All system parameters were found to be within specification except for the systematic error over ocean. With careful calibration scheme and detailed data retrieval algorithms accompanied with error correction algorithms, the errors can be reduced, still not perfectly removed. [28, p. 1356-1357]

An example time series of the SMOS brightness temperature data measured in the pixel where FMI-ARC is located is shown in Figure 2.7. As can be seen from the data, the variation at both polarizations is large and seasonal characteristics not that clear. Still, during winter periods the difference between polarizations is smaller than during summer, as it was in ELBARA-II data shown earlier in Figure 2.4. A differing feature is that only the horizontal polarization shows seasonal variation with larger values during winter, while such a feature is not visible in the vertical polarization.

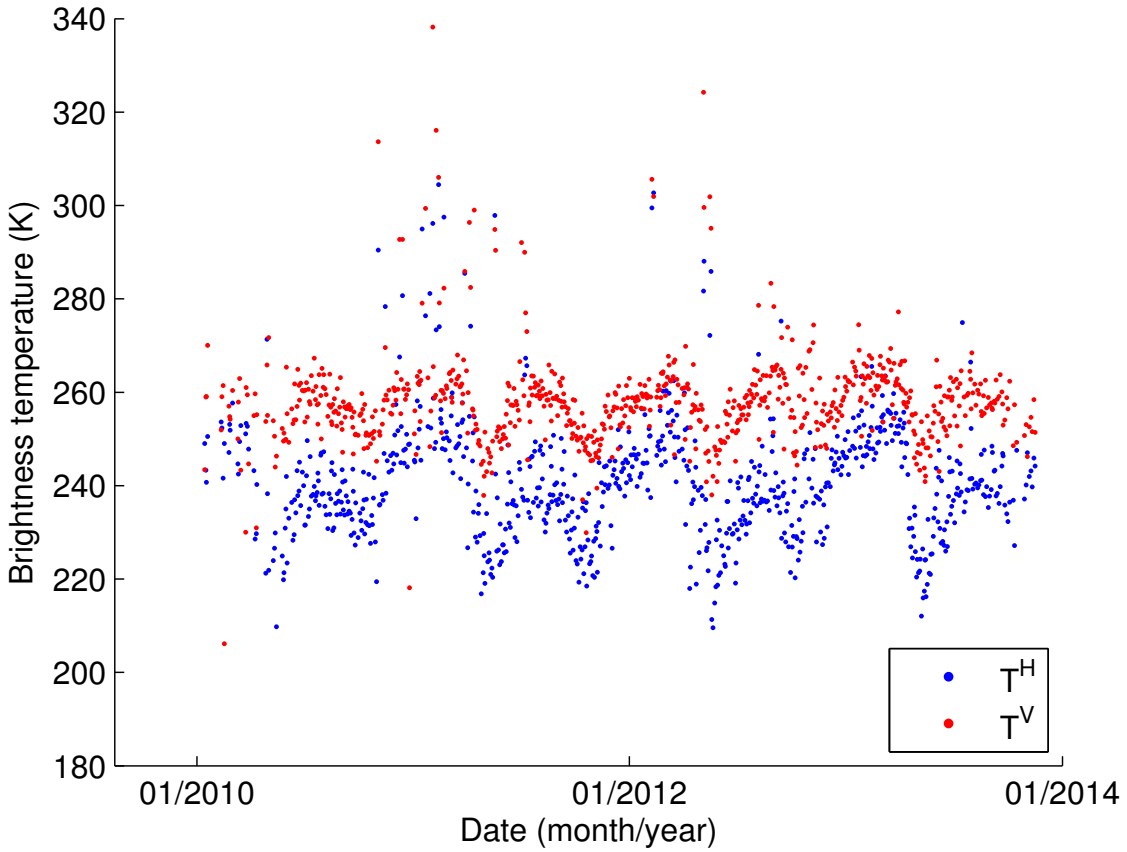

Figure 2.7 *Data measured by the SMOS satellite from the pixel where FMI-ARC is located with 50°-55° measurement angle. Outliers have not been removed.*

There are a few limitations in the concept chosen for the SMOS mission and in remote sensing methods in general. A significant caveat of remote sensing of soil moisture is that the penetration depth reaches only the surface layer of the soil, yet it is necessary to know the water amount available in the vadose zone as well. This could be achieved directly by using even lower frequencies leading to larger pixels, or indirectly by using assimilation methods whose performance is linked to the input data quality and the performance of the models used. Another limitation is the spatial resolution, which is still too coarse for many hydrological applications. Even though SMOS is operating in a protected band, the possibility of RFI is a serious issue. Also the effect of vegetation, land topography, possible snow cover and frozen soils can produce wrong estimates of the soil moisture. [19, p. 678].

As no space-borne L-band measurements had been made before the launch of the SMOS satellite, there were no similar data or previously used algorithms available for the SMOS calibration and validation purposes. For this reason, the extensive use of ground and aircraft data with simulations of the surface emission were chosen to validate the measurement methods used. As a part of this two core sites have

been equipped, manned and monitored throughout the SMOS mission and many campaigns and additional sites are being monitored continuously to check stability and seasonal variations of the signal [19, p. 680]. An important part of the validation process has also been the construction and the use of three identical L-band radiometers, the ELBARA-II instruments, described in more detail earlier in Section 2.3.1.

3. METHODS

Two different methods are tested for the estimation of the thermal state of the soil from the measured brightness temperature data. With the Canny edge detector algorithm, the dates of seasonal changes occurring in the brightness temperature data measured by the ELBARA-II radiometer and the SMOS satellite are searched. Secondly, regression analysis is applied to the ELBARA-II brightness temperatures and frost tube observations to gain estimates on the frost and thaw depths. In this chapter the theory and the assumptions behind the Canny edge detector algorithm and the multiple linear regression analysis are described in detail. Also the Kalman filter used for filtering the ELBARA-II data is introduced in this chapter.

3.1 Canny edge detector

In computer vision systems it is essential to identify features in images that are relevant when estimating the structure and properties of the objects. One such a feature are edges. The process of edge detection simplifies the analysis of images by reducing the amount of data to be processed, but it can also be applied for one dimensional edge profiles. Edges are significant local changes in the image intensity, typically occurring on the boundary between two regions in the image. Several methods have been proposed with differing filters, operators and thresholding procedures. For noisy edges, researchers have developed different optimal operators to overcome the unsatisfactory results obtained by using classical methods of gradient estimations. Some edge profiles and a noisy signal featuring such an edge are shown in Figure 3.1.

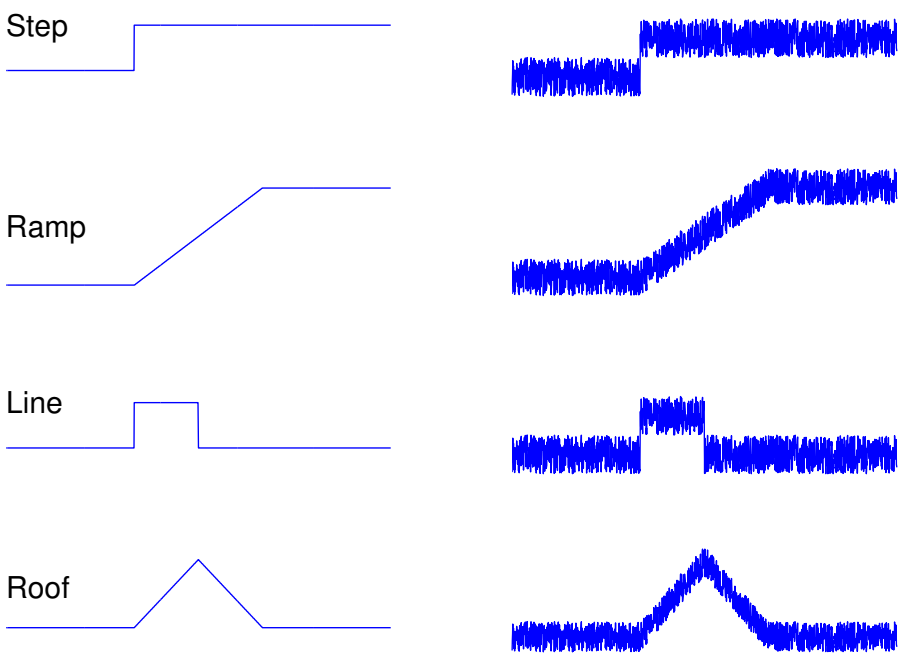

Figure 3.1 *Ideal one-dimensional edge profiles on the left and corresponding signals with additive white Gaussian noise on the right.*

In the design process of an edge detector the criteria used for the edge detector performance needs to be defined and mathematical formulations given. The first giving these in an analytical expression was John F. Canny in 1986. He defined three criteria, namely low error rate, good localization and only one response to single edge, which needed to be optimized in the edge detection. The first criteria emphasizes the importance of detecting the edges that occur in the image should not be missed and that only existent edges are detected. The localization criteria refers to the distance between the edges detected by the algorithm and the true edges which should be minimized. The last criteria is making sure that only one detector response is gained for a single edge.

Let's limit our interest on a one-dimensional edge detection problem where the goal is to find and mark step changes in a signal with additive white noise, like the one shown previously in Figure 3.1. It is assumed that the detection is performed by convolving the noisy edge with a continuous, finite filter function $f(x)$ and by marking the edges occurring at the maxima of this convolution. We aim to find a filter function $f(x)$ that is optimal in fulfilling the three criteria given previously.

Furthermore, it is assumed that the amplitude of the step at $x = 0$ is A, the noise $n(x)$ has variance n_0^2, the input signal is flat on both sides of the discontinuity and

that there are no other edges close-by. Now, the input signal $I(x)$ can be represented as

$$I(x) = Au_{-1}(x) + n(x),$$

where $u_{-1}(x)$ is the unit step function defined as

$$u_{-1}(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1, & \text{for } x \geq 0 \end{cases}.$$

Then, the output $H(x_0)$ when we apply convolution of I and f is

$$H(x_0) = (I * f)(x_0) = \int I(x)f(x_0 - x)\mathrm{d}x.$$

From the linearity of convolution it follows that the integral can be separated into the contributions due to the step only H_s and due to the noise only H_n. For a step occurring at $x_0 = 0$ the output $H(x_0)$ is

$$\begin{aligned} H(0) &= \int I(x)f(-x)\mathrm{d}x = \int \big(Au_{-1}(x) + n(x)\big)f(-x)\mathrm{d}x \\ &= \int Au_{-1}(x)f(-x)\mathrm{d}x + \int n(x)f(-x)\mathrm{d}x \\ &= \int Au_{-1}(-x)f(x)\mathrm{d}x + \int n(-x)f(x)\mathrm{d}x \\ &= A\int_{-\infty}^{0} f(x)\mathrm{d}x + \int n(-x)f(x)\mathrm{d}x \\ &= H_s(0) + H_n(0). \end{aligned} \tag{3.1}$$

From this, Canny defined the first criterion, the output signal-to-noise ratio SNR, as the quotient of the response to the step only and the root-mean-squared noise response, written as

$$\mathrm{SNR} = \frac{|H_s(0)|}{\sqrt{\mathrm{E}\left(H_n^2(0)\right)}} = \frac{A\left|\int_{-\infty}^{0} f(x)\mathrm{d}x\right|}{\sqrt{\mathrm{E}\left(\left(\int n(-x)f(x)\mathrm{d}x\right)^2\right)}}, \tag{3.2}$$

where $\mathrm{E}(y)$ is the expectation value of y. This can be written with an input dependent part and a filter dependent part as the following proposition states.

Proposition 1. *Let A be the amplitude of a noisy step occurring at $x = 0$. Let the noise $n(x)$ be white Gaussian noise with variance n_0^2. Then, the output signal-to-noise ratio,* SNR*, can be written as*

$$\mathrm{SNR} = \frac{A}{n_0}\Sigma, \tag{3.3}$$

where we have an input dependent part A/n_0 and a filter dependent part Σ defined as

$$\Sigma = \frac{|\int_{-\infty}^{0} f(x)\mathrm{d}x|}{\sqrt{\int f^2(x)\mathrm{d}x}}. \tag{3.4}$$

Proof. Assume the signal-to-noise ratio is defined from the signal output as the quotient of the response to the step only and the root-mean-squared noise response with

$$\mathrm{SNR} = \frac{A|\int_{-\infty}^{0} f(x)\mathrm{d}x|}{\sqrt{\mathrm{E}\left(\left(\int n(-x)f(x)\mathrm{d}x\right)^2\right)}}$$

as in (3.2). As the noise is assumed to be white, each of its components are statistically independent and the autocovariance of the noise $n(t)$, denoted with $\mathrm{C}_n(\tau)$, is of the form

$$\mathrm{C}_n(\tau) = q(t)\delta(t-\tau),$$

where $q(t) = \mathrm{E}\left(|n(t)|^2\right)$ at time t. White noise is stationary and is assumed to have zero mean, thus we have

$$q(t) = q = \mathrm{E}\left(|n(t)^2|\right) = \mathrm{V}\left(|n(t)|\right) - \left(\mathrm{E}(|n(t)|)\right)^2 = \mathrm{V}\left(|n(t)|\right) = n_0^2,$$

where $\mathrm{V}(y)$ is the variance of y. With the mean being zero, we also have

$$\mathrm{R}_n(\tau) = \mathrm{C}_n(\tau),$$

where $\mathrm{R}_n(\tau)$ is the autocorrelation of the noise $n(x)$.

Now, a result by [33, p. 312-313] states that if the input $x(t)$ to a linear system $y(t) = x(t) * h(t)$ is white noise with autocorrelation of the form $\mathrm{R}_x(\tau) = q(t)\delta(t-\tau)$, then

$$\mathrm{E}\left(y^2(t)\right) = q(t) * |h(t)|^2 = \int q(t-\alpha)|h(\alpha)|^2\mathrm{d}\alpha. \tag{3.5}$$

By using the result in (3.5), we can simplify the denominator of SNR to

$$\sqrt{\mathrm{E}\left(\left(\int n(-x)f(x)\mathrm{d}x\right)^2\right)} = \sqrt{\int n_0^2|f(x)|^2\mathrm{d}x} = n_0\sqrt{\int f^2(x)\mathrm{d}x}.$$

With this substitution to the denominator of SNR, we have

$$\mathrm{SNR} = \frac{A|\int_{-\infty}^{0} f(x)\mathrm{d}x|}{n_0\sqrt{\int f^2(x)\mathrm{d}x}}.$$

Defining Σ as

$$\Sigma = \frac{|\int_{-\infty}^{0} f(x)\mathrm{d}x|}{\sqrt{\int f^2(x)\mathrm{d}x}}$$

we arrive at (3.3). □

Now, finding the filter $f(x)$ which maximizes Σ in (3.4) corresponds to finding the best operator for detection only.

In the detection procedure the edges occurring in the signal were chosen to be marked at the maxima of the convolution of I and f. Since the input signal $I(x)$ contains noise, the maximum is expected to be displaced from the true position of the edge. The reciprocal of the root-mean-squared distance of the marked edge from the centre of the true edge serves as a measure which increases as localization improves. Since a maximum in the output function $H(x_0)$ can be found from the zero-crossing in the derivative of that output, we can write

$$H'(x_0) = \frac{\mathrm{d}}{\mathrm{d}x_0}\int f(x)I(x_0 - x)\mathrm{d}x = 0,$$

where the position of the edge x_0 we wish to find. Assuming the filter function $f(x)$ and its derivative $f'(x)$ are absolutely integrable, the convolution is differentiable [14, p. 387], and reduces to

$$\int f'(x)I(x_0 - x)\mathrm{d}x = 0.$$

In order to find x_0, we split the derivative of the output $H'(x_0)$ into components due to the step only and due to the noise only as done previously in (3.1), and we get

$$H'(x_0) = H'_s(x_0) + H'_n(x_0) = 0 \iff H'_s(x_0) = -H'_n(x_0), \tag{3.6}$$

from where it follows that

$$\mathrm{E}\left(H_s'^2(x_0)\right) = \mathrm{E}\left(H_n'^2(x_0)\right).$$

For the step only component $H_s'(x_0)$ we can write

$$H_s'(x_0) = \int f'(x) A u_{-1}(x_0 - x)\mathrm{d}x = \int_{-\infty}^{x_0} A f'(x)\mathrm{d}x = A f(x_0). \tag{3.7}$$

We can write the Taylor expansion of $H_s'(x_0)$ about the origin, which yields

$$H_s'(x_0) = H_s'(0) + H_s''(0)x_0 + O(x_0^2), \tag{3.8}$$

where the terms of higher order have been omitted by using the Landau's symbol $O(x_0^2)$ which describes the limiting behavior of a function. In the absence of noise there should be a local maximum at $x = 0$ in the response since the edges are centered at that point as assumed. So the first term in (3.8) can be ignored. When the displacement of the actual maximum from the true maximum is small, quadratic and all higher terms can be ignored as well. Combining this approximation with the result of (3.7) we get

$$H_s'(x_0) \approx H_s''(0)x_0 = A f'(0)x_0. \tag{3.9}$$

For the filter function f we need to add a further constraint that is should be antisymmetric, since if f has any symmetric component, the convolution with the input $I(x)$ will contribute to the noise component in the denominator of SNR in (3.3), resulting in a worse SNR than a purely antisymmetric filter [2, p. 681].

For the noise only component $H_n'(x)$, the response will be a Gaussian random vector with zero mean and variance equal to the mean-squared output amplitude and can be written

$$\mathrm{E}\left(H_n'^2(x)\right) = n_0^2 \int f'^2(x)\mathrm{d}x. \tag{3.10}$$

Now, combining (3.6) and (3.9), and using the result (3.10), we obtain the following expression for the expectation value of x_0^2

$$\mathrm{E}\left(x_0^2\right) \approx \mathrm{E}\left(\frac{H_n'^2(x_0)}{A^2 f'^2(0)}\right) = \frac{\mathrm{E}\left(H_n'^2(x_0)\right)}{A^2 f'^2(0)} = \frac{n_0^2 \int f'^2(x)\mathrm{d}x}{A^2 f'^2(0)}\ . \tag{3.11}$$

The localization criterion L is defined as the reciprocal of the square root of (3.11)

$$\mathrm{L} = \sqrt{\frac{A^2 f'^2(0)}{n_0^2 \int f'^2(x)\mathrm{d}x}} = \frac{A|f'(0)|}{n_0 \sqrt{\int f'^2(x)\mathrm{d}x}}. \tag{3.12}$$

Again, similarly as for the detection criteria, a measure Λ which is independent of the input signal $I(x)$ can be defined from (3.12) as

$$\Lambda = \frac{|f'(0)|}{\sqrt{\int f'^2(x)\mathrm{d}x}} \ . \tag{3.13}$$

Up to this point, two of the original three criteria have been defined in a mathematical form in (3.4) and (3.13). Both of these need to be maximized simultaneously and can be obtained by maximizing the product of these two, that is

$$\Sigma\Lambda = \frac{|\int_{-\infty}^{0} f(x)\mathrm{d}x|}{\sqrt{\int f^2(x)\mathrm{d}x}} \frac{|f'(0)|}{\sqrt{\int f'^2(x)\mathrm{d}x}}. \tag{3.14}$$

By spatial scaling of $f(x)$ it is possible to trade off detection performance against localization, but both cannot be improved simultaneously. Thus, there is an uncertainty principle for detection of noisy step edges [2, p. 684].

The analytic form of the first criterion in (3.4) was derived from the response at a single point at the centre of the edge and it does not take into account the possibility of several maxima nearby the edge center. Therefore, Canny added a third criterion, which ensures that there should be only one response into a single step edge in the vicinity of the step. This is achieved by limiting the number of peaks in the response, resulting in a lower probability of declaring more than one edge.

In an ideal case, the distance between peaks in the noise response approximate the width of the response of the filter operator to a single step, which is some fraction of the filter width, denoted with W [2, p. 682]. Since both minima and maxima of the output H produce zero-crossings in the derivative of the filter output, the mean distance between adjacent maxima in the output H is twice the distance between adjacent zero-crossing in the derivative of the filter output. In [39, p. 54], the statistical properties of random noise currents are discussed, and an expression for the expected number of zero-crossing per second passing a point is found. By

taking the reciprocal of the expression, we obtain the mean distance between the zero-crossings. For the response of a function g to Gaussian noise, this is

$$x_{\text{ave}} = \pi \left(\frac{-\,\mathrm{R}_g(0)}{\mathrm{R}''_g(0)} \right)^{\frac{1}{2}}.$$

The second derivative of the autocorrelation at τ can be written in a form [1, p.]

$$\mathrm{R}''_g(\tau) = -\int g'(t+\tau) g'(t) \mathrm{d}t.$$

Using the definition of autocorrelation and the previous result, $\mathrm{R}_g(0)$ and $\mathrm{R}''_g(0)$ can be written as

$$\mathrm{R}_g(0) = \int g^2(x) \mathrm{d}x$$

and

$$\mathrm{R}''_g(0) = -\int g'^2(x) \mathrm{d}x.$$

Now, by replacing the function g with the function f' for which we want the mean zero-crossing distance x_{zc}, we get

$$x_{\text{zc}} = \pi \left(\frac{\int f'^2(x) \mathrm{d}x}{\int f''^2(x) \mathrm{d}x} \right)^{\frac{1}{2}}. \tag{3.15}$$

As stated earlier, the distance $x_{\max}$ between adjacent maxima in the noise response of filter f is twice the distance between zero-crossings of f' defined in (3.15). When it is set to be some fraction k of the operator width W, we have

$$x_{\max}(f) = 2x_{\text{zc}}(f) = 2\pi \left(\frac{\int f'^2(x) \mathrm{d}x}{\int f''^2(x) \mathrm{d}x} \right)^{\frac{1}{2}} = kW. \tag{3.16}$$

This multiple response criteria is invariant to spatial scaling of f for any fixed value of k [2, p. 682,684].

As all the constraints have now been formulated, the edge detection problem reduces to the maximization of the product $\Sigma\Lambda$ in (3.14) under the constraint of the third criterion in (3.16). By using the calculus of variations, Canny found a linear, continuous filter for step edges that maximized the three performance criteria at any scale. The shape of the filter and the scale at which the edges can be found will depend on the choice of the operator width W. Since there is a trade-off between noise suppression and localization, reducing the noise will add uncertainty to the location

of the edge, on the contrary if the operator has greater sensitivity to detect edges it has increased sensitivity to noise also.

The best compromise between noise reduction and localization can be obtained by using the first derivative of a Gaussian as the filter. This corresponds to smoothing the data with a Gaussian filter and then computing the derivative of the output [35, p. 168-169]. Canny arrived to the same approximation for his optimal filter. The filter function $f(x)$ is now given by

$$f(x) = -\frac{x}{\sigma^2} \exp^{-\frac{x^2}{2\sigma^2}} .$$

Canny chose this approximation since there are efficient ways to compute the two-dimensional extension of the filter when it is represented as a derivative of a Gaussian. The filter function $f(x)$ and the output $H(x)$ when it is applied to a noisy edge signal $I(x)$ is shown in Figure 3.2.

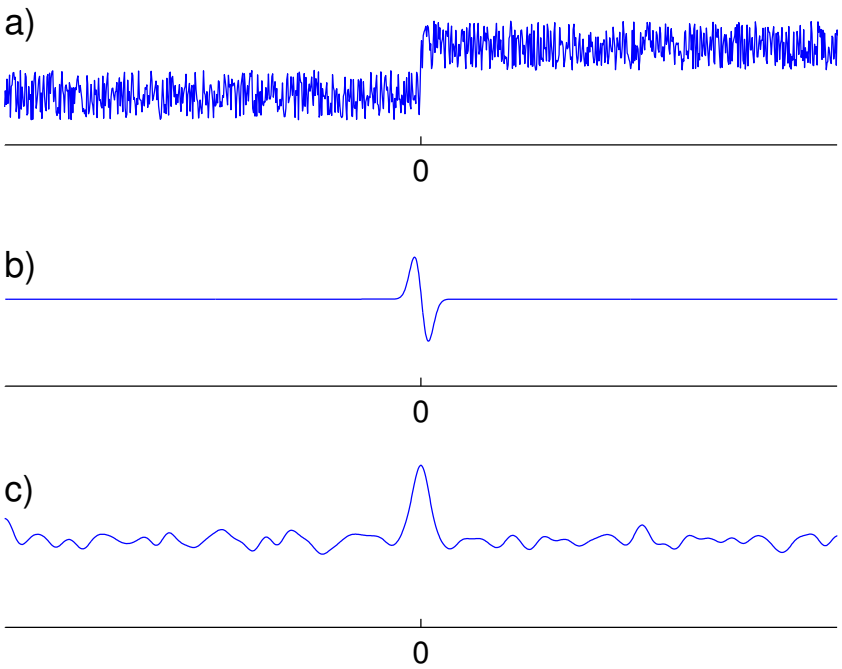

Figure 3.2 *a) The noisy edge signal $I(x)$ with ideal step edge. b) The filter function $f(x)$ which is the first derivative of a Gaussian. c) The output $H(x)$ when the filter function $f(x)$ is applied to the noisy edge signal $I(x)$.*

As can be seen from Figure 3.2, there are several maxima in the output $H(x)$, but there is one distinct maxima with a larger amplitude. This maxima clearly corresponds to the edge at $x = 0$ in the input signal $I(x)$. However, when applied to noisy real data with non-ideal step edge, there are more maxima in the output

$H(x)$ with larger amplitudes, making it more difficult to set apart the maxima corresponding to the true edge. In Figure 3.3 the filter function $f(x)$ is applied to a sample of the ELBARA-II brightness temperature measured at vertical polarization and and at 50° measurement angle. Also the output $H(x)$ is shown.

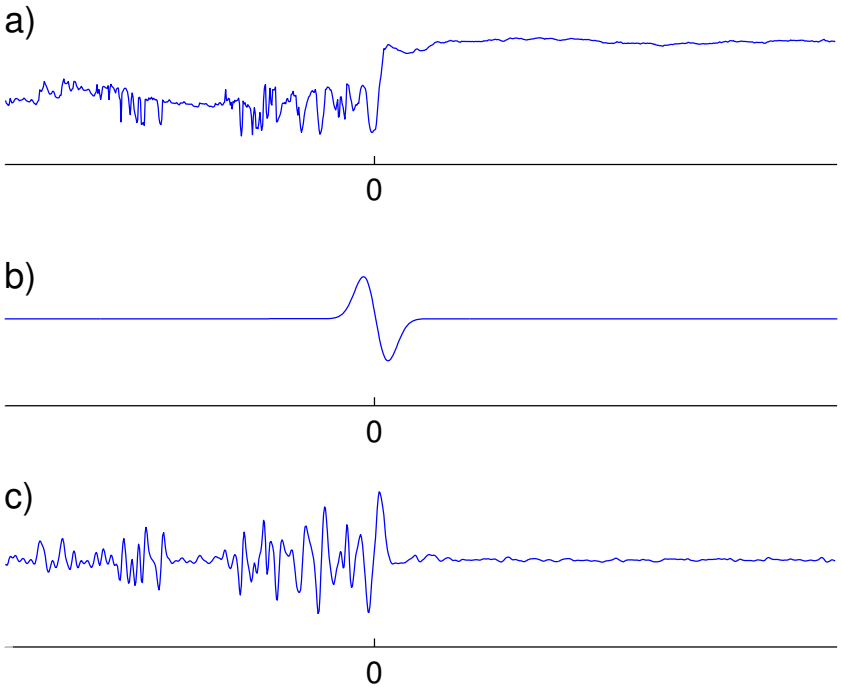

Figure 3.3 *a) The noisy edge signal $I(x)$, which is a sample from the ELBARA-II brightness temperature measured at vertical polarization and at 50° measurement angle. b) The filter function $f(x)$ which is the first derivative of a Gaussian. c) The output $H(x)$ when the filter function $f(x)$ is applied to a noisy edge signal $I(x)$.*

To find the edges at the maxima of the first derivative, a method of non-maximum suppression is applied where the magnitude of the derivative is compared between neighboring points. This is done to reduce the number of falsely detected edges. Here, Canny introduced the two threshold method, hysteresis, to overcome an issue with breaking up of the edges. In the presence of noise there will be fluctuation in the output amplitude above and below the threshold if the mean value of the operator corresponds the threshold chosen. This results in multiple falsely detected edges around one true edge. As pointed out by Canny, this is a pathological case. Nevertheless, choosing the thresholds so that the probability of marking noise edges as true edges would be small while retaining high sensitivity, is rather difficult. Noise estimation can be used when choosing the thresholds. [2, p. 689-690]

In the first part of hysteresis thresholding, all derivative magnitudes above a high threshold are marked as edges. Secondly, the maxima are compared to the lower

threshold and all the maxima above this are edges if they are also connected to an edge marked earlier. By using this procedure, the probability of breaking up of edges is reduced as the edge now needs to fluctuate above the high threshold and below the low threshold to be broken [2, p. 690]. This fine-to-coarse heuristic is chosen since the smallest operator gives the best localization. Yet, a trade-off between error rate and localization remains since choosing a high threshold leads to fewer false edges detected but the localization will be poorer since fewer edges are marked by the smaller operators [2, p. 692].

The feature synthesis procedure can be applied repeatedly beginning from small scale and advancing to larger scales to mark edges that were not marked by the previous scales. The majority of edges are marked by the smallest operator, while the latter mark mostly weaker edges. [2, p. 629]

Lastly, the Canny edge detection algorithm is presented in a compact form below.

Algorithm 1. *The Canny edge detector in one-dimension*

1. Filter the image by convolving the input signal $I(x)$ with a Gaussian filter $f(x)$ to filter out any noise.

$$H(x) = (I * f)(x)$$

2. Compute the derivative and determine the magnitude of change $G(x)$ in the image.

$$G(x) = \left| \frac{\mathrm{d}}{\mathrm{d}x} H(x) \right|$$

3. Find peaks of change by applying non-maximum suppression to the derivative magnitude to determine if the point is a better candidate for an edge than its neighbors.

$$G(x_i) = \begin{cases} 0, & \text{if } G(x_i) < G(x_{i+1}) \text{ or } G(x_i) < G(x_{i-1}) \\ G(x_i), & \text{otherwise} \end{cases}$$

4. Detect and link edges using hysteresis thresholding with low and high thresholds T_1 and T_2.

$$G(x) = \begin{cases} \text{definitely an edge} & \text{if } G(x) \geq T_2 > T_1 \\ \text{edge when neighboring point is an edge} & \text{if } T_1 \geq G(x) < T_2 \\ \text{definitely not an edge} & \text{if } G(x) < T_1 \end{cases}$$

3.2 Kalman filtering

Kalman filter is a recursive algorithm that produces estimates of unknown variables by using a series of noisy measurements observed over time. By taking into account all the previous and present measurements, the estimates obtained are more precise than those based on a single measurement. The filter is named after Rudolf E. Kálmán, one of the developers of the Kalman filtering theory. The algorithm works in a two-step process. In the first prediction step, the Kalman filter produces estimates of the current state, along with their uncertainties. Once we can use the next noisy measurement, these estimates are updated in the second step using a weighted average, with more weight being given to estimates with higher certainty. Because the algorithm is recursive, it can be executed in real time using only the present measurements and the previously calculated state and its uncertainty matrix. Since 1960 when Kálmán first introduced the Kalman filter in [17], several extensions and generalizations have been developed which work also on nonlinear systems. In this context we limit our interest on a linear system and errors with Gaussian noise.

To be able to estimate a state x by using the measurements y, we need a state model describing how the state x_k at a time t_k depends on the previous state at a time t_{k-1}. The state model is written as

$$x_k = G_{k-1}x_{k-1} + w_{k-1},$$

where $x_k \in \mathbb{R}^n$ is the state at a time t_k, $k \in \mathbb{N} \setminus \{0\}$, $G_{k-1} \in \mathbb{R}^{n \times n}$ is the state transition matrix and w_{k-1} is the state model error. The error is assumed to follow normal distribution with zero mean and a covariance Q_{k-1}. The matrix Q_{k-1} is assumed to be positive definite. [41, p. 56]

When the state is interpreted to be stochastic, the calculation of the state estimate $\hat{x}_k$ is done in two phases. In the first phase, the conditional distribution of x_k given the measurement y_k is solved. This conditional distribution is then used in the second phase to calculate the wanted weighted least square estimate [46, p.67]. At a time t_k, a measurement y_k of the true state x_k is made according to the measurement model

$$y_k = F_k x_k + v_k,$$

where $F_k \in \mathbb{R}^{d \times n}$ is the measurement model matrix and v_k is the measurement error assumed to follow normal distribution with zero mean and a covariance R_k. The matrix R_k is assumed to be positive definite. In addition to the state and measure-

ment models, information on the initial state x_0 is needed. Assume the initial state x_0 has a mean $\hat{x}_0$ and a covariance P_0. When the errors w_{k-1}, v_k and the initial state x_0 are assumed to be mutually independent, the conditional distribution of x_k given the measurements $y_1, y_2, \cdots, y_{k-1}$ can be solved. This conditional distribution is then used to obtain the parameters for the conditional distribution of x_k given the measurements $y_1, y_2, \cdots, y_k$, called the posterior. When the assumptions hold, the posterior follows normal distribution and its parameters can be solved by using a recursive algorithm, called the Kalman filter. [41, p.56]

The Kalman filter algorithm is presented in a compact form below. The proof of the algorithm can be found from [41, p.57-58]. For a more comprehensive derivation of the Kalman filter, see for example [26] or [41].

Algorithm 2. *The Kalman filter*

State model: $x_k = G_{k-1}x_{k-1} + w_{k-1}, \qquad \mathrm{V}(w_{k-1}) = Q_{k-1}.$
Measurement model: $y_k = F_k x_k + v_k, \qquad \mathrm{V}(v_k) = R_k.$
Measumerents: $y_{1:m} = \{y_1, y_2, ..., y_m\}$.
Initial state: estimate $\hat{x}_0$ *and covariance matrix* P_0.

1. Set k = 1.

2. Solve parameters for posterior distribution:

$$\begin{aligned} \hat{x}_k^- &= G_{k-1}\hat{x}_{k-1}, \\ P_k^- &= G_{k-1}P_{k-1}G_{k-1}^T + Q_{k-1}, \\ K_k &= P_k^- F_k^T \left(F_k P_k^- F_k^T + R_k\right)^{-1}, \\ \hat{x}_k &= \hat{x}_k^- + K_k\left(y_k - F_k\hat{x}_k^-\right) \end{aligned}$$

and

$$P_k = \left(I - K_k F_k\right) P_k^-.$$

3. Stop, if k = m, otherwise set k = k+1 and return to step 2.

3.3 Multiple linear regression analysis

The ultimate goal of regression analysis is to find relationships between a dependent variable and one or more independent variables. Many techniques have been developed for carrying out the analysis, including parametric methods such as linear regression and ordinary least squares regression, as well as nonparametric methods where the predictor is constructed according to information derived from the data. All these techniques include a number of assumptions made about the dependent and independent variables and the relationship between them, affecting the performance of regression analysis and the results obtained. In some cases these assumptions can be tested to verify that the fitted model actually fits the data and describes the relationship in a meaningful way.

In this context we limit our interest only on multiple linear regression analysis. The multiple linear regression model, the underlying assumptions made and the estimation of coefficients are introduced. Also the selection of the model and the variables, the tests for the goodness of the fit and for the statistical significance, as well as residual analysis are discussed in more detail in this section.

3.3.1 Multiple linear regression model

As mentioned above, the focus in regression analysis is on finding relationships between a dependent variable y, called the response, and one or more independent variables x, also called the explanatory variables. In order to predict or explain the behavior of the response y, a model can be constructed in the form of algebraic equation. The simplest regression model with linear relationship between the response y and one explanatory variable x_1 is the simple linear regression model, which can be written as

$$y = \beta_0 + \beta_1 x_1 + \epsilon, \tag{3.17}$$

where y is the response variable, β_0 and β_1 are the regression coefficients, x_1 is the explanatory variable and ϵ is a stochastic error term. The error ϵ is usually assumed to be independently and normally distributed with zero mean and variance σ^2, denoted with $\epsilon \sim N(0, \sigma^2)$. The regression coefficients β_0 and β_1 can be interpreted as the y-axis intercept point and the slope of the regression line, respectively.

With a straightforward extension the simple linear regression model can be formulated to include more than one explanatory variables. The definition of this multiple linear regression model is given next.

Definition 1. *Let y and $x_1, \ldots, x_k$ be random variables, $\beta_0, \ldots, \beta_k \in \mathbb{R}$ and ϵ an independent random variable that follows normal distribution with $\epsilon \sim N(0, \sigma^2)$. Assuming the relationship between the variables y and $x_1, \ldots, x_k$ is linear and that there is an additional error term ϵ, a multiple linear regression model describing the relationship can be written as*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon.$$

By defining $X \in \mathbb{R}^{m \times n}$, where $n = k + 1$, $y \in \mathbb{R}^{m \times 1}, e \in \mathbb{R}^{m \times 1}$ and $b \in \mathbb{R}^{k \times 1}$ as

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \ldots & x_{mk} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, e = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}, \text{ and } b = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

where x_{ij} represents the ith observation of the jth independent variable, $i = 1, \ldots, m$ and $j = 1, \ldots, k$, the multiple linear regression model can be expressed in matrix form as

$$y = Xb + e. \tag{3.18}$$

The linearity of the model refers to the model being linear with respect to its coefficients b. The explanatory variables $x_1, \ldots, x_k$ may depend on one another and for example polynomial models are linear with respect to its coefficients b, but not with respect to the independent variables.

In the simple linear regression model defined in (3.17), the regression coefficients β_0 and β_1 were interpreted as the y-axis intercept point and the slope of the regression line. For the multiple linear regression model, the interpretation of regression coefficients b becomes rather complicated since the estimates obtained are dependent on the other variables included in the model. A common but controversial interpretation for β_i is that it is the expected change in y for a unit change in x_i when all the other variables are held fixed. Here an assumption that changing one variable does not have any impact on the values of the other independent variables has been made. This does not hold if the variables are correlated. It might also be impossible to held a variable fixed. Thus it is relatively easy to misinterpret the regression results.

3.3.2 Estimation of coefficients

The objective is to find estimates such that the model defined in (3.18) best fits the observed data. As y and X are known, the task reduces to estimation of the unknown parameters b and σ^2. The estimates are denoted with $\hat{b}$ and $\hat{\sigma}$ respectively. Once these estimates are found, a fit $\hat{y}$ for y can be obtained.

A large number of procedures to obtain these estimates have been developed with varying computational efficiency, robustness and underlying assumptions needed. The simplest and most commonly used method for the estimation of b is the method of least squares which results in an estimate for which the sum of squared deviations from the response values y is minimized. We choose the vector b so that the sum $S(b)$ in

$$S(b) = \|y - Xb\|^2 = \sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k})^2 \tag{3.19}$$

is minimized, in which case we obtain the parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, which are gathered into a vector $\hat{b}$. The sum $S(b)$ in (3.19) can be written in matrix form as

$$\begin{aligned} S(b) &= \|y - Xb\|^2 = (y - Xb)^T(y - Xb) \\ &= y^T y - b^T X^T y - y^T Xb + b^T X^T Xb \\ &= y^T y - b^T X^T y - (b^T X^T y)^T + b^T X^T Xb \\ &= y^T y - 2b^T X^T y + b^T X^T Xb, \end{aligned} \tag{3.20}$$

where the last form is obtained by noticing that $y^T Xb \in \mathbb{R}^{1\times 1}$. The sum $S(b)$ is minimized when the partial derivatives of $S(b)$ with respect to the parameters $\beta_0, \beta_1, \cdots, \beta_k$ equal to zero. The estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ can then be solved from this set of equations, called the set of normal equations. In matrix form, we obtain this by differentiating (3.20) with respect to b and equating to zero which yields

$$(X^T X)\hat{b} = X^T y \iff \hat{b} = (X^T X)^{-1} X^T y.$$

Here it is assumed that X has full rank in which case $X^T X$ is positive definite and the inverse of $X^T X$ exists. The matrix $H = (X^T X)^{-1} X^T$ is often called the hat matrix because the fitted values $\hat{y}$ can be obtained by $\hat{y} = Hy$. The estimate $\hat{b}$ obtained is called the ordinary least squares estimate of b. The estimate $\hat{b}$ is an unbiased estimate of b, i.e. the difference between the estimate's expected value

and the true value of the parameter being estimated equals zero, as the following proposition states.

Proposition 2. *Let X be an $(m \times n)$ -dimensional matrix with rank n. If $\mathrm{E}(\epsilon) = 0$, then*

$$\hat{b} = (X^T X)^{-1} X^T y$$

is an unbiased estimate of b.

Proof. Assume the errors are unbiased, i.e. $\mathrm{E}(\epsilon) = 0$. Then

$$\mathrm{E}(\hat{b}) = \mathrm{E}\left((X^T X)^{-1} X^T y\right) = (X^T X)^{-1} X^T \mathrm{E}(y) = (X^T X)^{-1} X^T X b = b.$$

□

The estimate $\hat{b}$, which is also the maximum likelihood estimate of b, proves to have the minimum variance for the entire class of unbiased estimates [44, p. 43]. In maximum likelihood estimation the probability distribution of the random variable y is used to maximize the likelihood function. For more on maximum likelihood estimation see for example [34].

Having now obtained an estimate for b, an estimate for y, a fit $\hat{y}$, can be computed with

$$\hat{y} = X\hat{b}.$$

Now by defining the residual of y, denoted with $\hat{r}$, as

$$\hat{r} = y - \hat{y} = y - X\hat{b} = y - X(X^T X)^{-1} X^T y = \left(I - X(X^T X)^{-1} X^T\right) y, \qquad (3.21)$$

where I is the identity matrix, we can get a representation for that part of the response y that is not explained by the estimated model. In the literature the matrix $P = I - X(X^T X)^{-1} X^T = I - H$ is called the projection matrix, since P represents the orthogonal projection onto the column space of the matrix X. Using this notation the residual can be written as $\hat{r} = Py$.

The residual sum of squares, denoted with SS_{Res}, is the quantity

$$SS_{\mathrm{Res}} = \|\hat{r}\|^2 = \|y - \hat{y}\|^2 = \|y - X\hat{b}\|^2. \qquad (3.22)$$

By using SS_{Res}, an unbiased estimate for the error variance σ^2 can be obtained. It is given in the following proposition.

Proposition 3. *Let X be an $(m \times n)$ -dimensional matrix with rank n. If the expectation value and variance of the response y are $\mathrm{E}(y) = Xb$ and $\mathrm{V}(y) = \sigma^2 I$, respectively, then*

$$\hat{\sigma}^2 = \frac{\|y - X\hat{b}\|^2}{m - n} = \frac{SS_{\text{Res}}}{m - n} \tag{3.23}$$

is an unbiased estimate of σ^2

Proof. Omitted, see [44, p. 44-45] □

The estimate $\hat{\sigma}^2$ is commonly called the mean square error, denoted with MSE.

The model and the obtained parameter estimates and error variances can be used to compute the response in a new data point $\tilde{x}$. Assuming our multiple linear regression model (3.18) is valid at a new data point $\tilde{x}$, we can predict the corresponding response value z and compute the $(1-\alpha)$ confidence interval for it. There is a choice of either including or excluding the error term ϵ. When the error term is included, the true response z is $z = \tilde{x}b+\epsilon$, whereas the estimated response $\hat{z}$ is $\hat{z} = \tilde{x}\hat{b}+\epsilon$. Since $\epsilon \sim N(0, \sigma^2)$, $\mathrm{E}(z) = \tilde{x}b$ and $\mathrm{V}(z) = \sigma^2$. Because the components of e were assumed to be independent, $\hat{b}$ and e are independent, so z and $\hat{z}$ are also independent. Now $\mathrm{E}(\hat{z} - z) = \mathrm{E}(\hat{z}) - \mathrm{E}(z) = 0$ and $\mathrm{V}(\hat{z} - z) = \mathrm{V}(\hat{z}) + \mathrm{V}(z) = \sigma^2\tilde{x}^T(X^TX)^{-1}\tilde{x} + \sigma^2$ and we can compute the $(1 - \alpha)$ confidence interval limits for the response z with

$$\hat{z} \pm t_{\frac{\alpha}{2}}(m - n)\hat{\sigma}\sqrt{1 + \tilde{x}^T(X^TX)^{-1}\tilde{x}}. \tag{3.24}$$

Similarly, if the error term is excluded, the true response z is $z = \tilde{x}b$ and the estimated response $\hat{z}$ is $\hat{z} = \tilde{x}\hat{b}$. Now, $\mathrm{E}(\hat{z}) = z$ and $\mathrm{V}(\hat{z}) = \sigma^2\tilde{x}(X^TX)^{-1}\tilde{x}$ and the $(1 - \alpha)$ confidence interval limits for the response z can be computed with

$$\hat{z} \pm t_{\frac{\alpha}{2}}(m - n)\hat{\sigma}\sqrt{\tilde{x}^T(X^TX)^{-1}\tilde{x}}. \tag{3.25}$$

3.3.3 Hypothesis testing

A statistical hypothesis testing is a method of statistical inference to determine if a statement about the parameters describing a population holds true or not. As the

tests are based on random sampling, the test result can be interpreted as a random variable. Since the result is not definite, it can be incorrect, but the probability of this should be small and quantifiable.

The hypothesis testing starts by formulating two hypothesis, a null hypothesis H_0 and an alternative hypothesis H_1. The test is done assuming that the null hypothesis H_0 is true. The result either gives strong evidence to reject the null hypothesis H_0 and continue assuming that the alternative hypothesis H_1 is true, or does not give enough evidence to reject H_0 and the null hypothesis H_0 is still assumed to be true.

The result of hypothesis testing may be wrong in two ways. Type I error occurs when the null hypothesis H_0 is rejected although it is true. The probability of type I error is called the level of significance of the test, denoted with α. Type II error happens when H_0 is not rejected although is it false. Since the null hypothesis may be false in many ways, the probability of type II error cannot often be calculated.

In testing the null hypothesis for significance, a small probability, a significance level of the test α, is decided in advance to determine whether to reject the null hypothesis or not. Frequently used values are $\alpha = 0.05$ and $\alpha = 0.01$. If the test significance is equal or less than the α-value chosen, the probability of the outcome is sufficiently small and the null hypothesis is rejected.

For a parameter θ, there are three types of hypothesis tests, two one-tailed tests and a two-tailed test. The one-tailed hypothesis pairs are

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned} \tag{3.26}$$

and

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta < \theta_0 \end{aligned}, \tag{3.27}$$

and the two-tailed hypothesis pair is

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}, \tag{3.28}$$

where the value for θ_0 is given. The first pair in (3.26) is tested by calculating the lower $(1-\alpha)$ confidence limit $\hat{\theta}_L$ and the null hypothesis is rejected if the reference value θ_0 is not included in the obtained confidence interval. For (3.27), the test is done on computed upper $(1-\alpha)$ confidence limit $\hat{\theta}_U$ in a similar manner and for the

two-tailed hypothesis (3.28) the computed interval is a two-tailed $(1-\alpha)$ confidence interval $(\hat{\theta}_L, \hat{\theta}_U)$. Since the task is to verify if the value $\theta = \theta_0$ is included in the confidence interval or not, the testing does not require constructing the confidence interval itself.

The hypothesis testing is usually done by using test statistics. A test statistic is a measure of some attribute of a sample. Then, the null hypothesis is rejected if the computed value of the test statistic is in the critical region below the α-value chosen. The result of the hypothesis test can also be announced with a P-probability, called the p-value. The p-value is the smallest risk at which the null hypothesis H_0 can be rejected. Ideally the risk level α is chosen beforehand and the null hypothesis H_0 is rejected if the p-value is $\leq \alpha$.

In regression analysis, the model is considered to be insignificant if all the coefficients $\beta_1, \ldots, \beta_k$ are equal to zero. Correspondingly, if a coefficient β_i is zero, the explanatory variable x_i has insignificant effect on the response. In general, hypothesis testing is done on the obtained linear model against a reduced model where some of the variables are left out. For testing of general linear hypothesis, the hypothesis pair is

$$\begin{aligned} H_0 &: Kb = \kappa \\ H_1 &: Kb \neq \kappa \end{aligned},$$

where $K \in \mathbb{R}^{q \times n}$ and $\kappa \in \mathbb{R}^q$. It can be shown that when the null hypothesis is true, a test quantity F_t, which is a random variable, is F-distributed with q and $m-n$ degrees of freedom $F_t \sim F(q, m-n)$ [44, p. 100-101]. Defining the sum of squares of regression SS_{Reg} as

$$SS_{\text{Reg}} = \|\hat{y} - \bar{y}\|^2, \tag{3.29}$$

where $\bar{y}$ is the mean value of y, the test quantity F_t is obtained from

$$F_t = \frac{SS_{\text{Reg}}/q}{SS_{\text{Res}}/(m-n)}.$$

There is a connection between SS_{Res} defined in (3.22) and SS_{Reg} in (3.29), namely

$$SS_{\text{Tot}} = SS_{\text{Res}} + SS_{\text{Reg}},$$

where SS_{Tot} is called the total sum of squares. The corresponding mean squares are the mean square of regression MSR and the total mean square MST, respectively.

The test statistics computed can be gathered into a analysis of variance table, ANOVA-table, to help examining the regression results. Also comparison between fitted models becomes easier. An outline of the ANOVA-table is shown below in Table 3.1.

Table 3.1 *Outline of the ANOVA-table. DF stands for degrees of freedom.*

Source	DF	Sum of squares	Mean square	F-ratio	p-value
Model	$n-1$	SS_{Reg}	$\frac{SS_{Reg}}{n-1}$	$\frac{SS_{Reg}/(n-1)}{SS_{Res}/(m-n)}$	$\mathbb{P}(F > F_t)$
Error	$m-n$	SS_{Res}	$\hat{\sigma}^2$		
Total	$m-1$	SS_{Tot}			

Above, the testing was done on the whole model. For the testing of the significance of a single coefficient β_i, a test quantity t_i can be computed. The test quantity t_i follows the t-distribution with $m-n$ degrees of freedom $t_i \sim t(m-n)$ and it can be calculated by [44, p. 107]

$$t_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{a_i}},$$

where a_i is the $(i+1)$th diagonal element of the matrix $(X^T X)^{-1}$. When the corresponding p-value is smaller than the significance level α chosen, the test suggests to exclude the variable x_i from the model. These tests for different parameters are not independent. Thus, excluding several variables from the model may lead to unexpected results.

3.3.4 Selecting the model and the variables

The F-test described in earlier section can indicate that the formulated model is insignificant, in which case there is not much use for it. On the contrary, even if according to the F-test the model would be significant, the model might still not be a very good model. Possible reasons for this are, that the model might be unable to describe the behavior of the response, there might be some necessary explanatory variables missing from the model, or the model could be overspecified with too many explanatory variables. This is why selecting the variables to be included into the model is a very important part of the regression analysis.

The lack-of-fit-test can be used to test whether the model is adequate, i.e. if there are enough explanatory variables. If there are some necessary explanatory variables missing from the model, the model might not properly explain the variance of the

response. The null hypothesis H_0 is that the model is suitable. If it is rejected, there is a reason to revise the set of explanatory variables $x_1, \ldots, x_n$ and see, if more variables could be included into the model.

In general, the true error variance is not known and the lack-of-fit-test cannot be performed. In some cases it is possible to gain information of the true error variance from the data and the lack of fit of the model can be tested. One such a case occurs when there are multiple observations for one or more values of the explanatory variables. Let there be p groups of these multiple observations, each having q_i observations, where $i = 1, \ldots, p$. The variance s_i^2, where $i = 1, \ldots, p$ for each group is computed. Now, the combined variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{p}(q_i - 1)s_i^2}{\sum_{i=1}^{p}(q_i - 1)}$$

is an estimate for the true error variance. The residual sum of squares SS_{Res} can be separated into a sum of the "pure" variance and the variance caused by the lack of fit [5, p. 50], denoted as SS_{Pv} and SS_{Lf}. Let $q_v = \sum_{i=1}^{p}(q_i - 1)$ and $q_l = m - n - q_v$. It can be shown [44, p. 111-112], that if the model is adequate, the distribution of SS_{Lf} is a χ^2-distribution and the null hypothesis H_0 can be tested with a test quantity F_{Lf}

$$F_{\text{Lf}} = \frac{SS_{\text{Lf}}/q_l}{SS_{\text{Pv}}/q_v},$$

which follows the F-distribution with q_l and q_v degrees of freedom $F_{\text{Lf}} \sim F(q_l, q_v)$.

The lack-of-fit-test is not a typically used test for determining the need for additional variables, since there might not be observations with identical values available. And even though such multiple observations were found, the values for the variables to be added into the model might not be identical.

A commonly used parameter to examine how much the model explains the variations in the response y is the coefficient of determination R^2 defined as

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Tot}}}.$$

A value of R^2 closer to 1 tells the model explains much of the variations of the response. Still, if the R^2-value is very close to 1, there might be a danger of over-specified model. On the other hand, if the model is significant, it might be very

useful even if the coefficient of determination R^2 would be small, say 0.1. To be able to compare models with differing sizes, adjusted R^2 can be computed with

$$R^2_{\text{ad}} = 1 - \frac{SS_{\text{Res}}/(m-n)}{SS_{\text{Tot}}(n-1)}.$$

As was mentioned above, if the R^2-value is very close to 1, the model might be overspecified with too many explanatory variables. Since the residual sum of squares gets smaller as new explanatory variables are added into the model, one might think that if we just keep adding variables, the model keeps getting better and better. However, when the model has too many explanatory variables, it starts to explain the error instead of the underlying relationship. Also the precision of estimates and predicted values are more likely to be reduced for an overspecified model [12, p. 238]. Additionally, the interpretation of a model that has fewer explanatory variables is easier, while the coefficient of determination is only slightly reduced.

With calculation of Mallow's C_q number the issue of overspecific models can be tackled. As was mentioned above, by adding more variables, the residual sum of squares gets smaller. If this was the criterion for the selection of the model, the model including all the variables would always be selected. Here, all the possible subsets with p variables are searched and only the best of these models of the size q, $1 \leq q \leq p$ are selected into a closer analysis. The Mallow's C_q number, defined as

$$C_q = \frac{SS_{\text{Res}}(q)}{\hat{\sigma}^2} + 2q - m,$$

where $SS_{\text{Res}}(q)$ is the residual sum of squares of a model with q variables and $\hat{\sigma}^2$ is the variance estimate computed from the original model, is a commonly used criteria. If the model fits well, the bias caused by the model with q explanatory variables will be small and the expected value for C_q will be close to q.

There is also a class of stepwise methods for choosing the variables into the model, including backward elimination, forward selection and stepwise regression. In backward elimination, the starting point is a model with all the possible variables and variables are removed one at a time. The variable whose removal increases the residual sum of squares the least is removed. In the forward selection there is only the constant term to begin with, and new variables are added one at a time. Here, the variable whose addition decreases the residual sum of squares the least is added into the model. In the stepwise regression, these steps are conducted by turns. This stepwise method does not necessarily produce the best model, and frequently the

model obtained fails when it is applied to a new set of data. For these and other reasons this method has received criticism, and it should be used with a caution. [44, p. 392,419]

So far we have assumed that the explanatory variables are independent and that the matrix X^TX is non-singular. A phenomenon where there exists strong correlations among the explanatory variables is called multicollinearity. Because of the correlation, the matrix X^TX is singular or nearly singular and some of the parameter estimate variances are large. Multicollinearity can also be suspected when the tests indicate that the regression model is very significant while the p-values of t-test for the single coefficients are much larger. Reasons for multicollinearity are incomplete data collection methods resulting in an artificially created correlation, or that there naturally exists correlation among the variables. Since multicollinearity does not violate the underlying assumptions, the statistics obtained from the regression analysis are valid. Instead, the problem is in the data itself, because it is inadequate to properly estimate the regression coefficients [12, p. 177-178]. The extent of multicollinearity can be examined by using tests such as variance inflation factor and variance proportions [12, 44]. Multicollinearity does not affect the fit of the model or the estimated residual variation, thus multicollinearity can be ignored if the interest is on estimation and prediction, assuming predictions are not made at points remote from the observed data [44, p. 261].

Polynomial models are multicollinear, since terms of second order and higher tend to be highly correlated. One method to reduce the effects of multicollinearity is the use of a centered and scaled model. This is described in more detail next.

3.3.5 Centering and scaling the independent variables

The effect of centering and scaling of the explanatory variables $x_1, \ldots, x_n$ can be useful in polynomial models where multicollinearity can be a serious problem. The centering and scaling reduces multicollinearity, while the estimate precision is increased and the calculations become more accurate when the singularity of the matrix X is reduced.

Suppose that instead of the model in (3.18), we center the explanatory variables and use the reparametrized model

$$y = \alpha_0 + \beta_1\tilde{x}_1 + \cdots + \beta_k\tilde{x}_k + \epsilon,$$

where

$$\alpha_0 = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_k\bar{x}_k$$
$$\text{and}$$
$$\tilde{x}_i = x_i - \bar{x}_i.$$

Collecting the parameters $\alpha_0, \beta_1, \ldots, \beta_k$ into a vector $a^T = [\alpha_0, \beta_1, \ldots, \beta_k] = \left[\alpha_0, b_c^T\right]$, the variables $\tilde{x}_1, \ldots, \tilde{x}_k$ into a matrix $\tilde{X}$ and denoting $X_c = \begin{bmatrix}\mathbf{1} & \tilde{X}\end{bmatrix}$, the reparametrized model can be written in matrix form as

$$y = X_c a + e.$$

Because the transformation between a and b is one-to-one, the least squares estimate of b_c does not change. Now, since $\tilde{X}^T\mathbf{1} = \mathbf{0}$, the estimate $\hat{a}$ is

$$\begin{aligned}\hat{a} = (X_c^T X_c)^{-1} X_c^T y &= \begin{bmatrix} m & \mathbf{0}^T \\ \mathbf{0} & \tilde{X}^T\tilde{X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}^T y \\ \tilde{X}^T y \end{bmatrix} \\ &= \begin{bmatrix} m^{-1} & \mathbf{0}^T \\ \mathbf{0} & (\tilde{X}^T\tilde{X})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}^T y \\ \tilde{X}^T y \end{bmatrix} = \begin{bmatrix} \bar{y} \\ (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y \end{bmatrix},\end{aligned} \tag{3.30}$$

and we arrive at estimates $\hat{\alpha}_0 = \bar{y}$ and $\hat{b}_c = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$. Centering and reparametrization do not change the fitted model $\hat{y}$ and the residuals of both of the uncentered and centered models are the same. [44, p. 70-71]

We can also scale the columns of $\tilde{X}$ so that they have unit length. Let $s_j^2 = \sum_{i=1}^m (x_{ij} - \bar{x}_j)^2$ and $x_{ij}^* = (x_{ij} - \bar{x}_j)/s_j$, i.e. each column in $\tilde{X}$ is divided by the variance of that column. Now the model becomes

$$y = \alpha_0 + \gamma_1 x_{i1}^* + \cdots + \gamma_k x_{ik}^*,$$

where $\gamma_j = \beta_j s_j$. Since the transformation is still one-to-one, the estimates are $\hat{\gamma}_j = \hat{\beta}_j s_j$ and $\hat{\alpha}_0 = \bar{y}$. Collecting the elements x_{ij}^* into a matrix X^*, $\gamma_1, \ldots, \gamma_k$ into a vector g and replacing $\tilde{X}$ in (3.30) with X^*, we have

$$\hat{g} = (X^{*T}X^*)^{-1}X^{*T}y.$$

We can notice, that the matrix $X^{*T}X^*$ is actually the correlation matrix of the independent variables. [44, p. 71]

3.3.6 Correlated errors

In the regression models given earlier, the random errors were assumed to be uncorrelated. A common problem especially in time series data where the observations are collected in successive time periods is the possibility of correlated errors. The time-dependencies can be in the form of seasonal trends or cycles, or dependence on the previous time periods. Time-dependent errors are usually modeled with linear models, where the lack of independence is measured with correlations. These time-dependent errors are called autocorrelated or serially correlated errors. The correlation can be positive, when the residual tends to be the same as the sign of the preceding residual, or negative, when the signs of the residuals tend to alternate.

Some problems occur if the principle of least squares is used when the errors are serially correlated, particularly it they are positively correlated. The coefficient estimates $\hat{b}$ are still unbiased, but they are no longer the best linear unbiased estimates. The MSE in (3.23) might underestimate the variance of the error terms, also the obtained confidence intervals in (3.24) and (3.25) and hypothesis testing using the F- and t-statistics are no longer strictly applicable. [12, p. 164-165]

An autoregressive model for the correlated errors can be formulated. In the model the error ϵ_t at a time period t is linearly dependent on the previous errors. The autoregressive error model can be written as

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \cdots + \delta_t,$$

where ϵ_{t-i} is the error of the ith previous period, ρ_i is the correlation between the tth and $(t-i)$th error and δ_i is independent normally distributed random variable with mean zero and variance σ^2. The special case where the error of time period t is correlated only with the error of period $(t-1)$ is called the first-order autoregressive model, written as

$$\epsilon_t = \rho \epsilon_{t-1} + \delta_t. \tag{3.31}$$

As was mentioned, the positively correlated errors tend to form long series of residuals with the same sign. By plotting the residuals against the time, this pattern can be searched for. This simple but approximate test is called the runs test.

The most popular test for serial correlation is called the Durbin-Watson test after Durbin and Watson [6, 7, 8]. Suppose the errors ϵ_i follow the first-order autoregressive model formulated in (3.31). Let D be defined as

$$D = \frac{\sum_{i=2}^{m}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{m}\epsilon_i^2}.$$

The computed value for D is used to determine whether to reject the null hypothesis H_0 that $\rho = 0$ in which case there is no serial correlation. When the null hypothesis H_0 is true, the distribution of D depends on the matrix X, but Durbin and Watson found out that $D_L < D < D_U$, where the distributions of lower and upper bounds D_L and D_U do not depend on the data matrix X [6, p. 416]. In [7, p. 161-162] Durbin and Watson proposed the following procedure to test for the serial correlation. Let d be the computed value for D, α be the sample size, and $d_{L\alpha}$ and $d_{U\alpha}$ be the lower and upper tail significance points for the sample size α. If $d < d_{L\alpha}$, reject the null hypothesis, if $d > d_{U\alpha}$ accept the null hypothesis, otherwise the test is inconclusive. To test for the negative correlation, the statistic $4 - D$ can be used as the quantity $4 - d$ can be treated as though is it the value of D to be used for testing of positive correlation. [44, p. 292-294]

One method to take autocorrelation into account is to redefine the model used. In the case of first-order autocorrelation, the response at any time period t is dependent on the response at previous time period. Assuming that the autocorrelation is close to unity, a logical choice would be to model the change of response from time period to time period, called the first differences. Now, the redefined dependent variable d is the difference

$$d_t = y_t - y_{t-1}.$$

This model differs from the original model in (3.18), and often the results are not comparable [12, p. 170]

The explanatory variables can also be dependent on its own previous values, especially in the case of time series data. There are different autoregressive models that can be used when modeling time series data. These will not be discussed here, but for example [22] provides an insight into these.

3.3.7 Residual analysis

One important part of statistical analysis is to visualize the data and the results obtained. One method to investigate the model is to look at the residuals defined

earlier in (3.21). By examining the residuals it is also possible to detect extreme observations, called outliers, which differ from the other observations in the data set. These outliers can be removed from the dataset and the analysis redone to see it they have an influence on the estimated coefficients. Yet, there might be unique observations which do not fit well to the model which still are not outliers. Another class of differing observations is called influential observations which are observations that cause the regression estimates to be different from what they would be if these observations were removed from the data set [12, p. 120].

In general, outliers have large residuals but the residuals cannot be compared as such as the variance of residuals differs. For this reason we introduce internally Studentized residuals r_i defined as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}, \tag{3.32}$$

where h_i is the ith diagonal element of the hat matrix H. The hat matrix diagonals h_i can be interpreted as a measure of influence [44, p. 269]. If the ith data point is an outlier, we can exclude it from the dataset and fit a new multiple linear regression model to the rest of the data. Now we get an externally Studentized residual

$$t_i = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{1-h_i}}.$$

The influential observations can be searched by using Cook's distance D_i. When the ith observation is left out from the data set and a new model is fitted to the remaining data, the corresponding parameter estimate vector is denoted with $\hat{b}_{-i}$. Cook's distance D_i measures how much the fitted values $\hat{y}$ change when the parameters from the original model and the modified model are used. The Cook's distance D_i is defined as

$$D_i = \frac{(\hat{b}_{-i} - \hat{b})^T (X^T X)(\hat{b}_{-i} - \hat{b})}{n\sigma^2}.$$

Using internally Studentized residuals defined in (3.32), and a result that [44, p. 268]

$$\hat{b} - \hat{b}_{-i} = \frac{(X^T X)^{-1} x_i e_i}{1-h_i},$$

the Cook's distance D_i can be written

$$\begin{aligned} D_i &= \frac{(\hat{b}_{-i} - \hat{b})^T (X^T X)(\hat{b}_{-i} - \hat{b})}{n\sigma^2} \\ &= \frac{\left[-(X^T X)^{-1} x_i e_i\right]^T (X^T X) \left[-(X^T X)^{-1} x_i e_i\right]}{n\sigma^2 (1-h_i)^2} \\ &= \frac{x_i^T (X^T X)^{-1} (X^T X)(X^T X)^{-1} x_i e_i^2}{n\sigma^2 (1-h_i)^2} = \frac{x_i^T (X^T X)^{-1} x_i e_i^2}{n\sigma^2 (1-h_i)^2} \\ &= \frac{h_i e_i^2}{n\sigma^2 (1-h_i)^2} = \frac{1}{n} \frac{e_i^2}{\sigma^2 (1-h_i)} \left(\frac{h_i}{1-h_i}\right) \\ &= \frac{1}{n} r_i^2 \left(\frac{h_i}{1-h_i}\right). \end{aligned}$$

The Cook's distance will be large if a point i has a large Studentized residual r_i or if it is a high-influence point with large h_i.

The results obtained from residual analysis can be plotted in diagnostic plots to see if the assumptions made are valid. The interpretation of the plots can be challenging if the amount of data is limited.

Yet another graphical tool to identify deviation from the assumption of normality and to detect outliers, is called the normal probability plot. It is formed by plotting ordered residuals against selected quantiles of the normal distribution. Deviations from a straight line suggest the assumption of normality might not be valid. Outliers can also be detected as isolated points at the plot, whereas skewed errors show up as a curve. [44, p. 295-296]

4. RESULTS

The Canny edge detector algorithm, Kalman filtering and multiple linear regression analysis described in Chapter 3 were applied to real data measured with the ELBARA-II radiometer and the SMOS satellite to detect the thermal state of the soil and to estimate frost depth. The results are given below in separate sections according to the method used. The calculations and visualization were conducted using MATLAB version R2013b.

4.1 Canny edge detector

The seasonal changes in the brightness temperature data measured by the ELBARA-II instrument and the SMOS satellite were searched using the Canny edge detector algorithm described in Algorithm 1. The results from ELBARA-II measurements are given first and the results from the SMOS data follow.

As can be seen from the time series of ELBARA-II data shown earlier in Figure 2.4, the brightness temperatures T_B measured by the ELBARA-II radiometer contain outliers at both polarizations. There are also some discrete missing observations due to maintain actions and electricity blackouts. For these reasons, the brightness temperature data was interpolated using shape-preserving piecewise cubic interpolation. The first year's measurements were omitted due to data quality issues and differing measurement interval.

In the high latitudes of Sodankylä where the ELBARA-II radiometer is located, the soil is frozen continuously for several months. Also the freezing and thawing periods are distinct and repetitive freeze-thaw cycles are rare. Therefore, a reasonable assumption is to locate only two edges per year in the brightness temperature data, one occurring in the autumn and another one in the spring, corresponding to freezing and thawing of the soil respectively. For this reason, the scale at which we are searching for edges is quite coarse and only one threshold, denoted with T, will be used. As was discussed in Section 3.1, the σ^2 value chosen for the Gaussian filter determines the degree of noise suppression and the scale at which the edges can be

found. Several combinations for the value pairs of σ^2 and T were tested. The results obtained for horizontal and vertical polarizations are shown in Figures A.1 and A.2.

As can be seen from the Figures A.1 and A.2, the smaller the scale and the threshold, the more there are edges marked. For brightness temperatures at horizontal polarization the best localization with the correct number of edges is obtained at five different combinations of σ^2 and T, those being when $\sigma^2 = 64$ and $T = 0.7$, $T = 0.8$ or $T = 0.9$, and when $\sigma^2 = 128$ and $T = 0.5$ or $T = 0.6$. From these only the dates corresponding to the edges found at the smallest thresholds are shown in Table 4.1, since the edges found on larger thresholds have exactly the same location. For brightness temperatures at vertical polarization, the best localization with the correct number of edges is obtained at two different combinations of σ^2 and T, namely when $\sigma^2 = 32$ and $T = 0.7$, and when $\sigma^2 = 64$ and $T = 0.7$. The dates corresponding to these edges are shown in Table 4.1 along with the difference to the starting dates of freezing or thawing periods located from the measured frost tube data.

Table 4.1 *The dates corresponding to the edges found from the horizontal and vertical ELBARA-II brightness temperature data. Also the difference between the dates and the starting date of the freezing or thawing periods located from the frost tube measurements are shown. Positive difference means, that the edge was found at a later datum than the starting date of the freezing or thawing period located from the frost tube measurements.*

H		V	
$\sigma^2 = 64, T = 0.7$		$\sigma^2 = 32, T = 0.7$	
05.11.2010	+20	09.11.2010	+24
02.05.2011	+6	06.05.2011	+10
09.11.2011	-7	11.11.2011	-5
16.05.2012	-5	17.05.2012	-4
$\sigma^2 = 128, T = 0.5$		$\sigma^2 = 64, T = 0.7$	
10.11.2010	+25	09.11.2010	+24
29.04.2011	+3	02.05.2011	+6
08.11.2011	-8	09.11.2011	-7
09.05.2012	-12	15.05.2012	-6

The edges found at the scales and thresholds shown in Table 4.1 correspond quite well with the starting and ending of frost period observed from the measured frost tube data. The difference in the autumn 2010 can be a result of the soil being frozen only from the top few centimetres in late October, while only a few week's later the

frost depth had reached over 30 cm. Otherwise, the difference between the dates of the edges found and the starting dates of freezing and thawing periods located from the frost tube data, is only in the scale of a few days'. And even though the edges occurring in the brightness temperature data are not ideal step edges as assumed in the Canny edge detector algorithm, the overall performance of the algorithm for ELBARA-II measurements with edges more of the shape of a ramp is good and the algorithm could be used to estimate the start of the freezing and thawing periods.

The Canny edge detector algorithm was tested also on the noisier SMOS brightness temperature data, which has also both lower temporal sampling and larger spatial resolution. Before applying the algorithm, obvious outliers were removed and the SMOS data was interpolated for daily brightness temperature values using shape-preserving piecewise cubic interpolation. Several combinations for the value pairs of σ^2 and T were tested. The results obtained for horizontal and vertical polarizations are shown in Figures A.3 and A.4.

As can be seen from the Figures A.3 and A.4, more edges are marked at smaller scales and thresholds as was from the ELBARA-II data, but a differing feature is that there are more edges found than expected and with a poorer localization. For brightness temperatures at horizontal polarization the most promising localization are obtained at three different combinations of σ^2 and T, those being when $\sigma^2 = 8$ and $T = 0.5$, when $\sigma^2 = 32$ and $T = 0.6$, and when $\sigma^2 = 64$ and $T = 0.4$, $T = 0.5$ or $T = 0.6$. From these only the dates corresponding to the edges found at the smallest thresholds are shown in Table 4.2, since the edges found on larger thresholds have exactly the same location. For brightness temperatures at vertical polarization, the most promising localization is obtained at two different combinations of σ^2 and T, namely when $\sigma^2 = 32$ and $T = 0.5$, and when $\sigma^2 = 64$ and $T = 0.6$. The dates corresponding to these edges are also shown in Table 4.2 along with the difference between the starting dates of freezing or thawing periods located from the measured frost tube data.

Table 4.2 *The dates corresponding to the edges found from the SMOS brightness temperature data. Also the difference between the dates and the starting dates of the freezing or thawing periods located from the frost tube measurements are shown. Positive difference means, that the edge was found at a later datum than the starting date of the freezing or thawing period located from the frost tube measurements. Differences over 42 days are shown in bold. The consecutive winters are separated with dashed lines.*

H				V	
$\sigma^2 = 8, T = 0.5$		$\sigma^2 = 64, T = 0.4$		$\sigma^2 = 32, T = 0.4$	
06.04.2010	-30	29.03.2010	-38	15.04.2010	-21
16.06.2010	+41	18.06.2010	+**43**	08.06.2010	+33
08.11.2010	+23	17.11.2010	+32	12.11.2010	+27
01.04.2011	-25	31.03.2011	-26	10.04.2011	-16
16.05.2011	+20	18.06.2011	+**53**	05.06.2011	+40
15.11.2011	-1	29.11.2011	+13	23.09.2011	-**54**
15.02.2012	-**96**	14.04.2012	-37	21.11.2011	+5
09.04.2012	-42	07.07.2012	+**47**	27.04.2012	-24
				09.06.2012	+19
				15.07.2012	+**55**
21.10.2012	-5	14.12.2012	+**49**	16.09.2012	-40
14.04.2013	-17	07.04.2013	-24	18.10.2012	-8
28.05.2013	+27	25.06.2013	+**55**	24.12.2012	+**59**
17.10.2013	+1			16.04.2013	-15
				28.05.2013	+27
$\sigma^2 = 32, T = 0.6$				$\sigma^2 = 64, T = 0.6$	
10.04.2010	-26			04.04.2010	-32
07.06.2010	+32			16.06.2010	+41
14.11.2010	+29			19.11.2010	+34
05.04.2011	-21			06.04.2011	-20
13.06.2011	+**48**			18.06.2011	+**53**
17.11.2011	+1			19.09.2011	-**58**
16.04.2012	-35			11.12.2011	+25
				17.04.2012	-34
				04.07.2012	+**44**
13.12.2012	+**48**			02.01.2013	+**68**
14.04.2013	-17			06.04.2013	-25
30.05.2013	+29			19.06.2013	+**49**

The edges found and shown in Table 4.2 do not correspond as well with the starting and ending of frost periods observed from the measured frost tube data as the edges found from the ELBARA-II data did. The correspondence is better in the autumn time than in the spring time, when in most cases one edge is marked approximately one month prior to the starting of the thawing period and another one approximately one month after. As the seasonal changes are not that clear in the SMOS brightness temperature data and it has lower temporal sampling and larger spatial resolution, the inaccuracy of the algorithm is understandable. Still, an overall feature that can be seen is, that when an edge corresponds better to the starting of the frost period, the edge is marked later than the actual frost period had started. In the spring time the effect is opposite as an edge is marked before the actual start of the thawing period. One possible explanation is that these edges might better reflect the average start of the frost and thaw periods in the Sodankylä pixel, where the dominant land type is peatland. In large bog areas, where there are more water stored into the soil than in the moorland, the frost period can start later because the energy stored into the water slows down the cooling process of the soil. In the thawing period, the effect of moist land is not that straightforward as the presence of liquid water in the soil surface and overlying snow make the radiative transfer much more complex than in the freezing period as discussed earlier in Subsection 2.3.1.

In addition to the data measured only over Sodankylä pixel, the performance of the Canny edge detector algorithm was tested on a larger scale for the interpolated and Kalman filtered SMOS brightness temperature data measured over Finland during autumn 2010. The edges were searched from both horizontal and vertical polarizations with several combinations for the value pairs of σ^2 and T. The dates of the edges found were categorized according to the corresponding week of the year. In Figure 4.1 one example of the results obtained for horizontal brightness temperature data when $\sigma^2 = 32$ and $T = 0.4$ is shown.

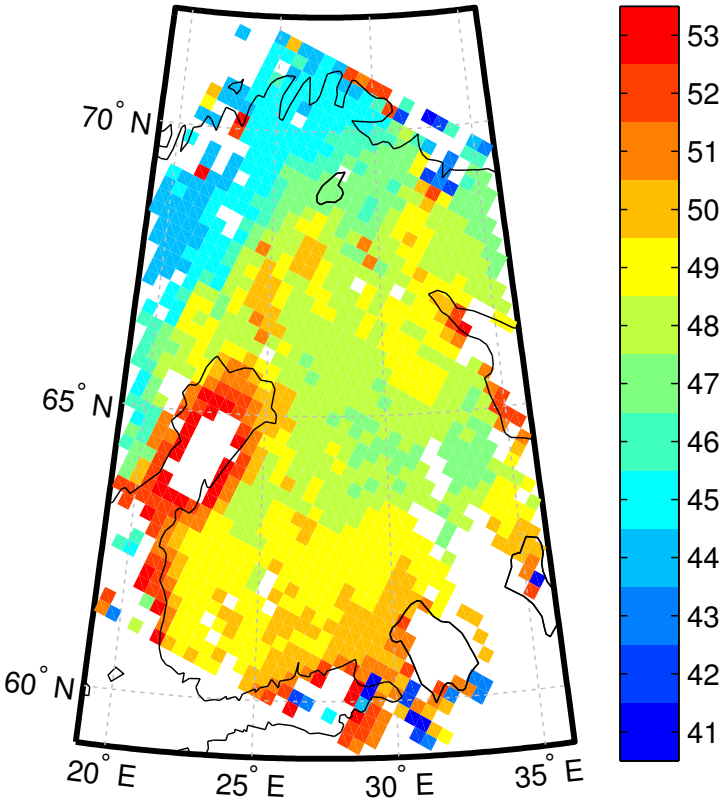

Figure 4.1 *The week of year when an edge was found from the horizontal brightness temperature measured by the SMOS satellite over Finland during autumn 2010. For white pixels there was no data available or no edges were found. The borders of land areas and large lakes are shown in black. The pixels over waters were not omitted from the figure.*

As can be seen from Figure 4.1, the soil appears to freeze earlier in the north than in the south as expected. Also the neighboring pixels show similar behavior as they should. However, there are many pixels over land where no edges were found. As these results were not compared to any in situ measurements or other reference dataset, we cannot say anything about the accuracy of the results obtained. Nevertheless, the results look very promising and the methodology could be improved and developed further.

Overall, even though the results from the Canny edge detector algorithm were not that accurate, it could be used as a tentative tool when searching for seasonal changes from a time series of brightness temperature data. One major caveat still remaining is that by using the Canny edge detector algorithm, one cannot gain any estimate on the current frost and thaw depth conditions, as the algorithm only marks the times when there are step changes in the brightness temperature data.

4.2 Regression analysis

To gain estimates on the frost and thaw depths from the brightness temperature data measured by the ELBARA-II instrument, an algorithm described in [38] was applied to the brightness temperature data. In addition to strictly following this methodology, where temporal averaging is applied to the data with a moving average filter, the Kalman filter described in Section 3.2 was also tested. Third and last approach was to obtain two simpler linear models, one for the freezing period and another one for the thawing period. These three approaches and the results obtained are discussed next.

In [38, p. 212], the terms raw frost factor and relative frost factor are introduced as the starting point for the calculation of the soil state estimate. Seven different formulations for the calculation of the raw frost factor are given. According to the sensitivity analysis performed to all of these factors, the frost factor denoted with FF_{combH}, and defined as

$$FF_{\mathrm{combH}} = (T_B^{\mathrm{V}} - T_B^{\mathrm{H}})(270 - T_B^{\mathrm{H}}) = 270T_B^{\mathrm{V}} - 270T_B^{\mathrm{H}} - T_B^{\mathrm{V}}T_B^{\mathrm{H}} + {T_B^{\mathrm{V}}}^2$$

was found to have the highest accuracy [38, p. 213]. Thus in this context we limit our interest only to this raw frost factor. The raw frost factor FF_{combH} is further used to compute the relative frost factor FF_{rel} defined as

$$FF_{\mathrm{rel}} = \frac{FF_{\mathrm{combH}} - FF_{\mathrm{SUMMER}}}{FF_{\mathrm{WINTER}} - FF_{\mathrm{SUMMER}}},$$

where FF_{WINTER} and FF_{SUMMER} are the mean winter and summer frost factors respectively. The definitions of FF_{WINTER} and FF_{SUMMER} used here differ from the definitions of [38], where FF_{WINTER} and FF_{SUMMER} were defined from the whole three year observation period as the average of 30 of the largest and smallest values of FF_{combH} with 33% of the most extreme values omitted [38, p. 212]. Here we use a procedure, where from each three freezing periods the average of minimum values and the average of maximum values are chosen to represent FF_{WINTER} and FF_{SUMMER} respectively.

As the emitted signal decays exponentially in the soil medium, an exponential model $f(x)$ is fitted into the observations in [38]. The model $f(x)$ is formulated as

$$FF_{\mathrm{rel}} = \beta_0(1 - \exp(-\beta_1 \cdot FD) + v = f(FD) + v, \tag{4.1}$$

where FD is the frost depth, β_0 and β_1 are coefficients and v is an additional error term not originally included in the model introduced in [38]. The error term v is

assumed to follow normal distribution with zero mean. The variance of the error term v is estimated to equal the maximum likelihood estimate of the error variance $\hat{\sigma}^2$ computed for the fitted exponential model [5, p. 507].

Once the fit is computed, the estimated frost depth $\widehat{FD}$ can be obtained by using a formula

$$\widehat{FD} = -\frac{\log(1 - \frac{FF_{\mathrm{rel}}}{\beta_0})}{\beta_1}. \tag{4.2}$$

As there are considerable diurnal variations in the measured brightness temperature data, temporal averaging was applied to the computed FF_{combH} values in [38]. Here we chose to apply temporal averaging to the computed FF_{rel} instead of FF_{combH} values and, in addition, to compare between two filters, the moving average filter used originally in [38] and the Kalman filter described in Section 3.2. The computed fits with computed 5% and 95% Bayesian confidence interval bounds [23, p. 36-37, 54] are shown in Figures 4.2 and 4.3.

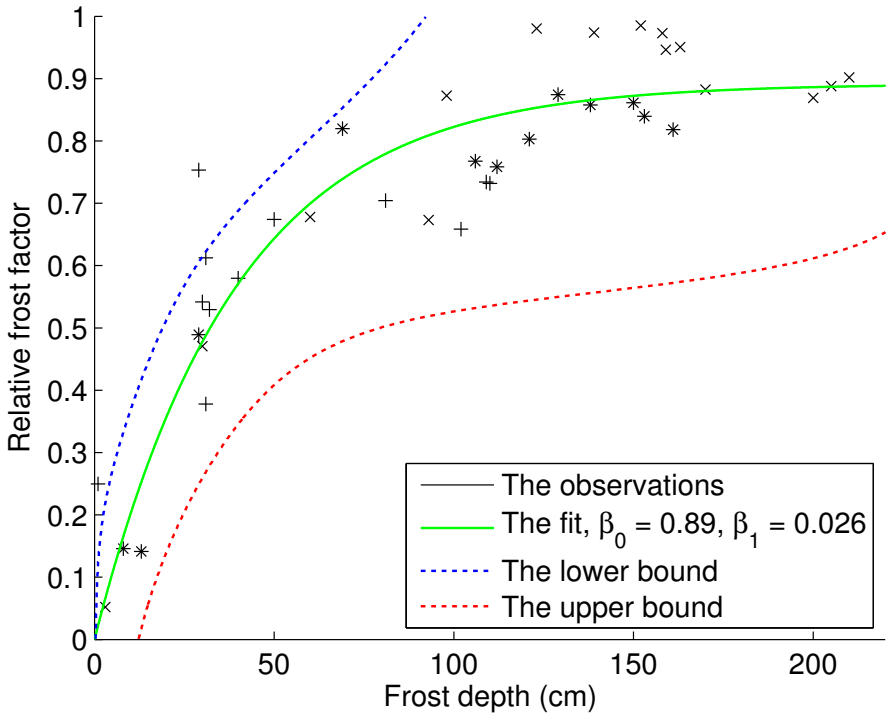

Figure 4.2 *Exponential model formulated in (4.1) fitted into the data filtered with the moving average filter. The observations from different winters are shown with differing black markers. Also the 5% and 95% Bayesian confidence interval bounds are shown.*

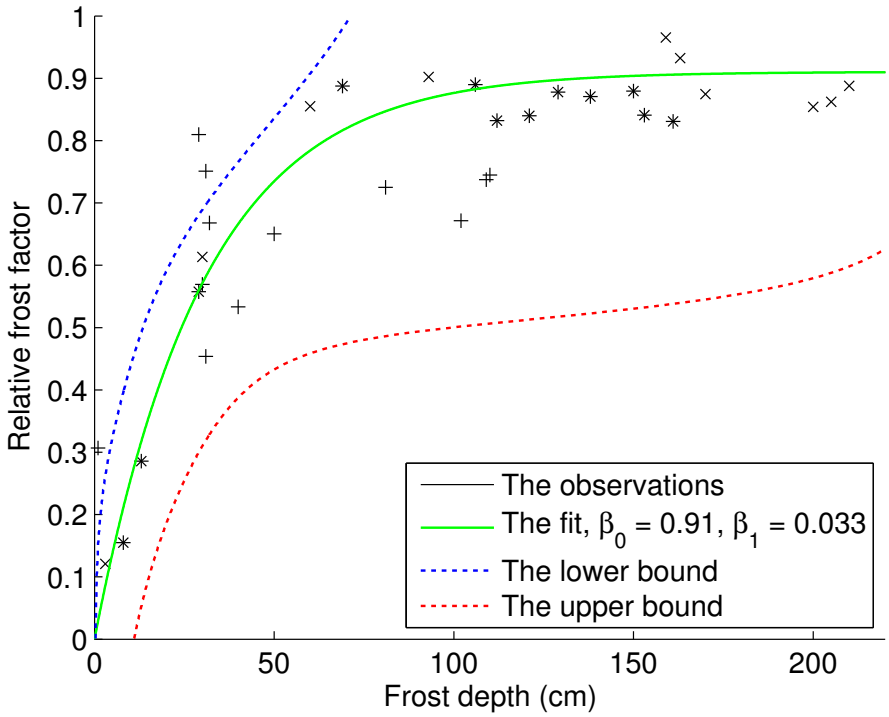

Figure 4.3 *Exponential model formulated in (4.1) fitted into the data filtered with the Kalman filter. The observations from different winters are shown with differing black markers. Also the 5% and 95% Bayesian confidence interval bounds are shown.*

Using these two fits, the estimated frost depths were calculated from the filtered FF_{rel} values using (4.2). The frost depths were further divided into three categories, those being "Unfrozen", "Freezing/Thawing started" and "Soil frozen". The definitions for the categories are given in Table 4.3. These definitions were chosen based on the accuracy of the categorized results when being compared to the frost tube measurements. The computed soil state estimates are shown in Figure B.1.

Table 4.3 *Definitions of the three selected soil state categories.*

Soil state category	Definition	Value
Unfrozen	$\mathbb{P}(FD < 2) > 0.95$	0
Freezing/Thawing started	$\mathbb{P}(2 \leq FD < 4) > 0.95$	-100
Soil frozen	$\mathbb{P}(FD \geq 4) > 0.95$	-200

As can be seen from Figure B.1, the soil state estimates computed from both the moving average filtered data and the Kalman filtered data show good correspondence with the observed frost depth. In the autumn 2009, there are large variations in the soil state estimates and both fail to properly follow the progressing of the freezing front. During the two other freezing periods the correspondence is better. On

the contrary, during the freezing period in the autumn 2011 the measured frost depth does not lie inside the computed 90 % Bayesian confidence interval, while it does during the other two freezing periods. In the spring times both the soil state estimates have difficulties in estimating the start of the thawing period, but the soil state estimate computed from the moving average filtered data changes in the vicinity of the starting of the thawing period. During both winter and summer periods there are no false categorizations in either of the soil state estimates. Overall, the accuracy of the soil state estimate computed from the moving average filtered data is smaller than the soil state estimate computed from the Kalman filtered data, since the accuracies of right categorizations are 75,8% and 87,8% respectively.

The third and last approach was to obtain separate linearized models for the freezing and thawing periods to see if a simpler model would be adequate to describe the freezing and thawing periods' processes. In order to be able to formulate such a linear model, let the variables $x_1, \ldots, x_5$ denote the horizontal brightness temperature, vertical brightness temperature, the product of horizontal and vertical brightness temperatures, and the second powers of horizontal and vertical brightness temperatures respectively. Now, a polynomial model with additional error term v can be formulated as

$$\begin{aligned} FD &= \beta_0 + \beta_1 T^H + \beta_2 T^V + \beta_3 T^H T^V + \beta_4 {T^H}^2 + \beta_5 {T^V}^2 + v \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + v \\ &= Xb + v. \end{aligned} \tag{4.3}$$

Once the fit is computed, the estimated frost depth $\widehat{FD}$ can be computed with

$$\widehat{FD} = X\hat{b}. \tag{4.4}$$

For the modeling of the freezing period, the frost depth values below 50 cm were selected from the frost tube measurements. Since the frost tubes are being read only at an interval of 10 days, the selected frost depths were interpolated to cover the same time span as the corresponding ELBARA-II measurements. The polynomial model given in (4.3) was then fitted by using three different sets of brightness temperature data, these being the raw data, the moving average filtered data and the Kalman filtered data. From these, the models fitted using the moving average filtered data and the Kalman filtered data showed very similar results and the choice on which model to use was done on the basis of the results obtained from the hypothesis testing and by examining the obtained R^2_{Ad} values. The model fitted using the

moving average filtered data showed better results and was thus selected. The ANOVA-table of this model is shown in Table 4.4.

Table 4.4 *The ANOVA-table for the model of the freezing period. DF stands for degrees of freedom.*

Source	DF	Sum of squares	Mean square	F-ratio	p-value
x_1	1	10715	10715	421.0	$2.27 \cdot 10^{-88}$
x_2	1	4213	4213	165.6	$4.85 \cdot 10^{-37}$
x_3	1	232	232	9.1	$2.56 \cdot 10^{-3}$
x_4	1	8648	8648	339.8	$1.92 \cdot 10^{-72}$
x_5	1	2554	2554	100.3	$2.66 \cdot 10^{-23}$
Error	3448	87751	25.5		
Total	3453	$4.9 \cdot 10^5$			

As can be seen from Table 4.4, all the variables are significant according to the p-values of the F-test. The fit of the model is good according to the computed R^2 and R^2_{Ad} values which are both 0.82. To detect possible outliers and to see if the assumptions made when formulating the model are valid, residual analysis was performed. The diagnostic plots and the case order plots of leverage and Cook's distance are shown in Figure 4.4.

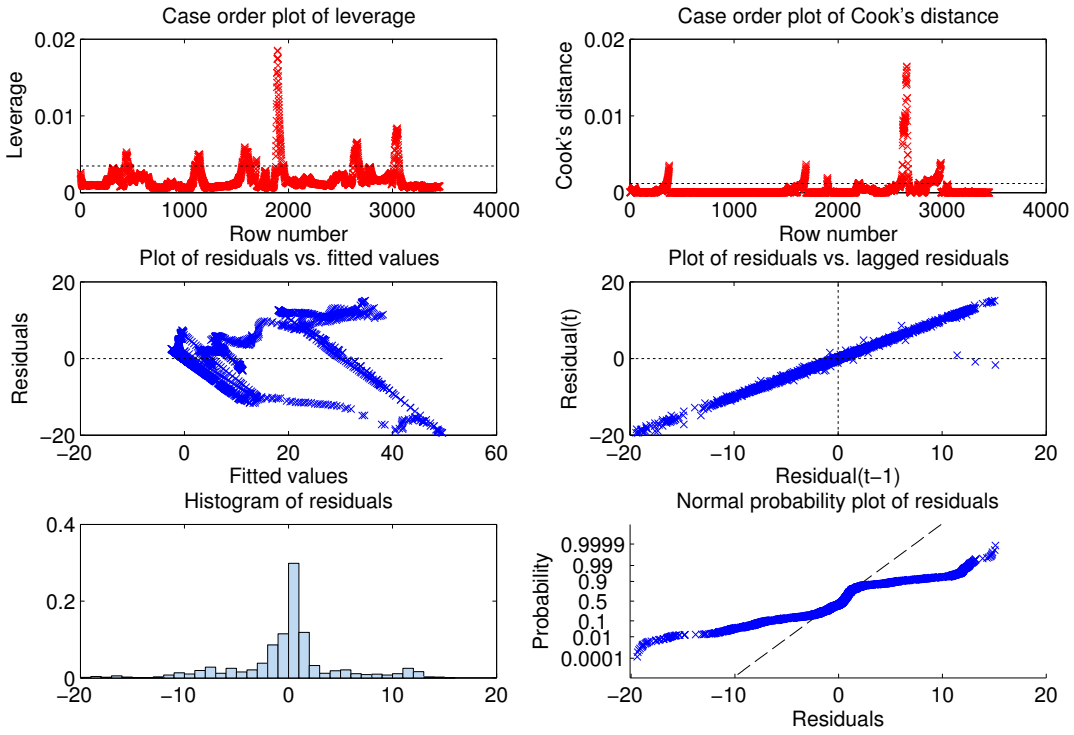

Figure 4.4 *The case order plots of leverage and Cook's distance and diagnostic plots for the linear model obtained for the freezing period.*

Both the case order plots shown in Figure 4.4 indicate that there might be some observations which are potential outliers and should be removed from the dataset. The histogram of residuals confirms this. The assumptions on the normality and uncorrelated errors seem to be invalid according to the normal probability plot and the plot of residuals versus lagged residuals. The latter was expected, since we are using the moving average filtered brightness temperature data resulting a brightness temperature value on a certain time to be dependent on the values of the previous times.

Since the residual analysis indicates that there are potential outliers, a new model was fitted when observations with large residuals were omitted. The ANOVA-table of this modified model is shown in Table 4.5.

Table 4.5 *The ANOVA-table for the model of the freezing period when observations with large residuals have been omitted. DF stands for degrees of freedom.*

Source	DF	Sum of squares	Mean square	F-ratio	p-value
x_1	1	8585	8595	499.8	$3.17 \cdot 10^{-103}$
x_2	1	3209	3209	186.6	$2.27 \cdot 10^{-41}$
x_3	1	91	91	5.3	0.02
x_4	1	15958	15958	927.9	$6.69 \cdot 10^{-180}$
x_5	1	1198	1198	69.7	$1.01 \cdot 10^{-16}$
Error	3314	56995	17.2		
Total	3319	$3.6 \cdot 10^5$			

When these statistics of the modified model are compared to the statistics of the original model shown in Table 4.4, nearly all the statistics show improved values. For the variable x_3 corresponding to the product of horizontal and vertical brightness temperatures, the p-value is now larger but the null hypothesis is not rejected when the significance level is $\alpha = 0.05$. The fit of the modified model is slightly improved as the computed R^2 and R^2_{Ad} values are both 0.825. The diagnostic plots and the case order plots of leverage and Cook's distance for the modified model are not shown here since they indicate no other possible outliers and the other features are very similar to those of the original model shown in Figure 4.4.

Following the procedure of model fitting for the freezing period, a model for the thawing period was also fitted. The only difference was the use of measured thaw depth values below 25 cm instead of frost depth values. The model fitted using the moving average filtered data showed the best results and was thus selected. The ANOVA-table of this model is shown in Table 4.6.

Table 4.6 *The ANOVA-table for the model of the thawing period. DF stands for degrees of freedom.*

Source	DF	Sum of squares	Mean square	F-ratio	p-value
x_1	1	400	400	28.7	$1.98 \cdot 10^{-7}$
x_2	1	671	671	48.1	$3.58 \cdot 10^{-11}$
x_3	1	115	115	8.2	$4.49 \cdot 10^{-3}$
x_4	1	37	37	2.6	0.11
x_5	1	419	419	30.0	$1.05 \cdot 10^{-7}$
Error	246	3431	13.9		
Total	351	12767			

The variable x_4 corresponding to the second power of the horizontal brightness temperature has large F-test p-value as can be seen from the Table 4.6. This indicates that it should be removed from the model when the significance level is set to $\alpha = 0.05$. The fit of the model is good as the computed R^2 and R^2_{ad} values are 0.731 and 0.726 respectively. To detect possible outliers and to see the validity of the assumptions made when formulating the model, residual analysis was performed before removing any variables from the model. The diagnostic plots and the case order plots of leverage and Cook's distance are shown in Figure 4.5.

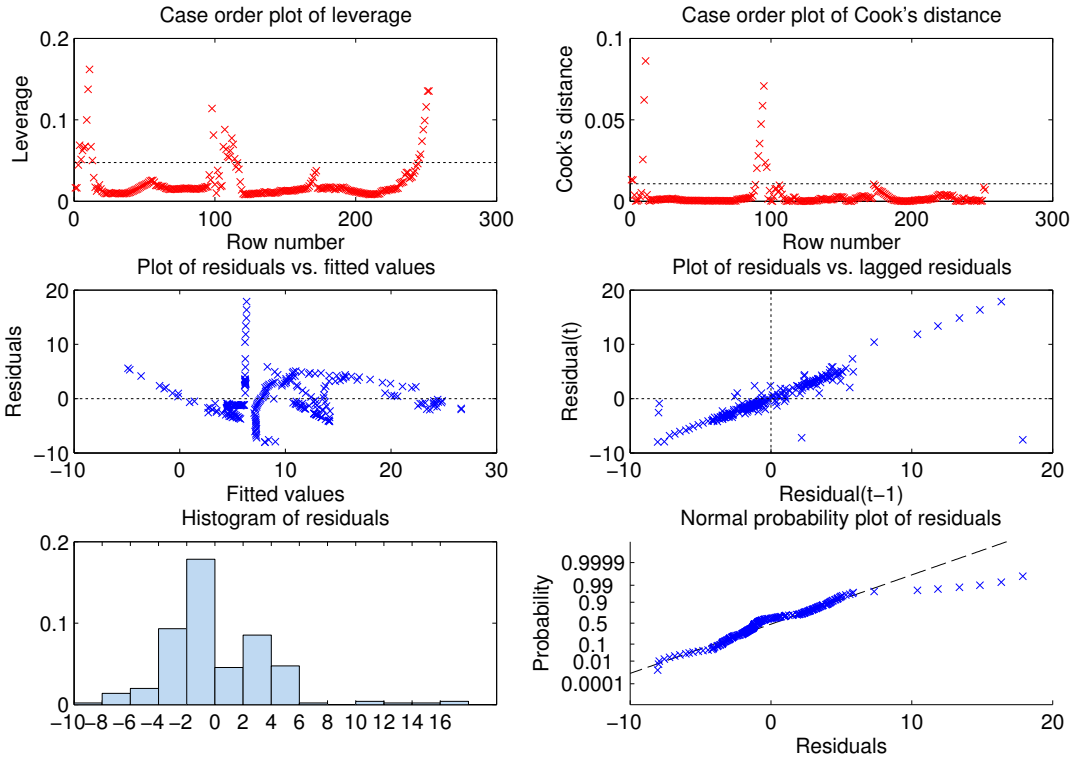

Figure 4.5 *The case order plots of leverage and Cook's distance and diagnostic plots for the linear model obtained for the thawing period.*

Both the case order plots shown in Figure 4.5 indicate that there might be some observations which are potential outliers and should be removed from the dataset. The histogram of residuals confirms this. The assumption on the uncorrelated errors seems to be invalid according to the plot of residuals versus lagged residuals. This was expected, since we are using the moving average filtered brightness temperature data resulting a brightness temperature value on a certain time to be dependent on the values of the previous times. There is also some deviation from normality as the normal probability plot shows.

Before removing any variables from the model, a new model was fitted when observations with large residuals were omitted. The ANOVA-table of this modified model is shown in Table 4.7.

Table 4.7 *The ANOVA-table for the model of the thawing period when observations with large residuals have been omitted. DF stands for degrees of freedom.*

Source	DF	Sum of squares	Mean square	F-ratio	p-value
x_1	1	338	338	44.7	$1.69 \cdot 10^{-10}$
x_2	1	508	508	67.1	$1.72 \cdot 10^{-14}$
x_3	1	75	75	9.9	$1.84 \cdot 10^{-3}$
x_4	1	42	42	5.5	$1.96 \cdot 10^{-2}$
x_5	1	310	310	41.0	$8.48 \cdot 10^{-10}$
Error	233	1763	7.6		
Total	238	11453			

When these statistics of the modified model are compared to the statistics of the original model shown in Table 4.6, nearly all the statistics show improved values. For the variable x_4 corresponding to the second power of horizontal brightness temperature, the p-value is now below the significance level $\alpha = 0.05$ and thus there is no need for removal of the variable. The fit of the modified model is better than the original model as the computed R^2 and R^2_{ad} values are 0.846 and 0.843 respectively. The normal probability plot for the modified model shows less deviations from the normality, but otherwise the features of the diagnostic plots and the case order plots of leverage and Cook's distance are similar to those of the original model shown in Figure 4.5 and thus these plots are not shown here.

Using these two fits obtained separately for the freezing and thawing periods, the estimated frost and thaw depths were calculated using (4.4). The frost depths calculated for the model fitted using freezing period were further divided into three

categories, those being "Unfrozen", "Freezing/Thawing started" and "Soil frozen". The definitions for the categories are given in Table 4.8. These definitions were chosen based on the accuracy of the categorized results when being compared to the frost tube measurements.

Table 4.8 *Definitions of the three selected soil state categories for the model fitted using the freezing periods.*

Soil state category	Definition	Value
Unfrozen	$FD < 15$	0
Freezing/Thawing started	$15 \leq FD < 20$	-100
Soil frozen	$FD \geq 20$	-200

For the thawing period model, the interpretation of the calculated thaw depths is rather complicated, since the thaw depths vary in the range of -88 to +80 cm. It seems still, that the largest thaw depth values appear in the vicinity of the start of the freezing and thawing periods. Thus, the soil state estimate was set to "Freezing/Thawing started" when the estimated thaw depth was 15 cm or more. The computed soil state estimates are shown in Figure B.2.

As can be seen from Figure B.2, the soil state estimate computed from the model fitted to the freezing period shows good correspondence with the observed frost depth. During the freezing period there is only one false categorization in January 2012. In the spring time there are difficulties in estimating the starting of the thawing period. During winter and summer periods there are no false categorizations and the overall performance is good as the percentage of right categorizations is 80.1%. The thawing period model manages to indicate the starts of the freezing and thawing periods quite well but it could be better as the percentage of right categorizations is only 47.3%.

5. CONCLUSION

Detection of the thermal state of the soil is essential since it has a great impact on the global water, energy and carbon cycles, which are important factors in numerical models of the atmosphere. With a good knowledge of the soil's thermal state the performance of atmospheric models and hydrological and climatological forecasts could be improved. In this master of science thesis the term thermal state of the soil was introduced and recent development in the field of retrieving the soil's thermal state with measurements performed at ground level and by remote sensing methods were discussed. Two instruments, the ELBARA-II radiometer locating in Sodankylä, Finland, in FMI-ARC and the SMOS satellite carrying an L-band MIRAS radiometer were discussed in more detail.

The brightness temperature data measured by the ELBARA-II radiometer and the SMOS satellite were used to estimate the thermal state of the soil with two different methods. The first method was the Canny edge detection algorithm with which the time of a seasonal change in the measured brightness temperatures was located. The Canny edge detector algorithm proved out to have a good performance when applied to the local ELBARA-II brightness temperature data and the changes detected corresponded well with the start of the freezing and thawing periods detected from in situ frost tube measurements. The performance and the accuracy were not as good for the SMOS brightness temperature data, which has both lower temporal sampling and larger spatial resolution. Yet, when the algorithm was applied to the SMOS brightness temperature data measured over Finland during autumn 2010, the results were promising as the neighboring pixels showed similar behavior and the soil freezing seemed to progress towards the south. Thus, the possibility of using the Canny edge detector algorithm as a tentative tool when searching for seasonal changes from a time series of brightness temperature data could be studied more. However, one major caveat remains as by using the Canny edge detector algorithm, one cannot gain any estimate on the current frost and thaw depth conditions, as the algorithm only marks the times when there are changes in the brightness temperature data.

The second method was the applying of regression analysis to the brightness temperature data measured by the ELBARA-II radiometer to gain estimates on the frost and thaw depths. Three different approaches were tested. In the first, an algorithm described in [38] was applied to the moving average filtered brightness temperature data. In the second, the same procedures were applied to Kalman filtered brightness temperature data. The third and last approach was to formulate two polynomial models, one for the freezing period and another one for the thawing period to see if these simpler models would be adequate to describe the freezing and thawing periods' processes. The frost and thaw depth estimates computed using these three models were further divided into three soil state categories. The best accuracy of the computed soil state estimates was obtained for the second approach, where the Kalman filtered brightness temperature data was used. The formulation of a simple thawing period model proved out to be challenging, but the simple freezing period model showed more accurate results than the first approach where the moving average filtered data was used.

The work presented in this master of science thesis could be extended in a number of ways. Firstly, more edge detector algorithms could be tested to see if these would give more accurate results, especially for the SMOS brightness temperature data. Also the methodology of searching edges from the SMOS brightness temperature data measured over larger land areas could be developed further to get more accurate results. Secondly, more research could be done on the use of regression analysis to estimate the frost and thaw depths. One interesting aspect would be to test the regression analysis on SMOS brightness temperature data. Also the use of data measured at different angles and from different pixels and different latitudes could be tested. The regression model could be modified to include a categorical variable to account for the land class of the pixel where the data was measured from. Also the use of an autoregressive model would be worth testing, as it might suit better for the analysis of time series data.

REFERENCES

[1] J. Bendat and A. Piersol, *Random data: Analysis and measurement procedures*, 4th ed. Wiley, 2010, 640 p.

[2] J. Canny, "A computational approach to edge detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[3] I. Corbella, F. Torres, A. Camps, A. Colliander, M. Martín-Neira, S. Ribó, K. Rautiainen, N. Duffo, and M. Vall-Ilossera, "MIRAS end-to-end calibration: application to SMOS L1 processor," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 5, pp. 1126–1134, 2005.

[4] *5TE Water Content, EC and Temperature Sensors*, 8th ed., Decagon Devices,Inc., 2007-2012.

[5] N. Draper and H. Smith, *Applied regression analysis*, 3rd ed. Wiley, 2014, 738 p.

[6] J. Durbin and G. Watson, "Testing for serial correlation in least squares regression. I," *Biometrika*, vol. 37, pp. 409–428, 1950.

[7] J. Durbin and G. Watson, "Testing for serial correlation in least squares regression. II," *Biometrika*, vol. 38, pp. 159–177, 1951.

[8] J. Durbin and G. Watson, "Testing for serial correlation in least squares regression. III," *Biometrika*, vol. 58, pp. 1–19, 1971.

[9] D. Entekhabi, E. Njoku, P. O'Neill, K. Kellogg, W. Crow, W. Edelstein, J. Entin, S. Goodman, T. Jackson, J. Johnson, J. Kimball, J. Piepmeier, R. Koster, N. Martin, K. McDonald, M. Moghaddam, S. Moran, R. Reichle, J. Shi, M. Spencer, S. Thurman, L. Tsang, and J. Van Zyl, "Soil moisture active passive (SMAP) mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[10] European Centre for Medium-Range Weather Forecasts (ECMWF) web site, available (accessed on 13.8.2014): http://www.ecmwf.int/.

[11] O. Frauenfeld, T. Zhang, and J. McCreight, "Northern hemisphere freezing/thawing index variations over the twentieth century," *International Journal of Climatology*, vol. 27, pp. 47–63, 2007.

[12] R. Freund, W. Wilson, and P. Sa, *Regression Analysis*, 2nd ed. Elsevier Science, 2006, 481 p.

[13] M. Hallikainen, F. Ulaby, M. Dobson, M. El-Rayes, and L.-K. Wu, "Microwave dielectric behavior of wet soil - Part I: Empirical models and experimental observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 23, no. 1, pp. 25–34, 1985.

[14] K. Howell, *Principles of Fourier analysis.* Taylor and Francis, 2001, 791 p.

[15] A. Ishimaru, *Wave propagation and scattering in random media*, 1st ed. Elsevier Science, 2013, 272 p.

[16] Y. Iwata, T. Hirota, T. Suzuki, and K. Kuwao, "Comparison of soil frost and thaw depths measured using frost tubes and other methods," *Cold Regions Science and Technology*, vol. 71, no. 1, pp. 111–117, 2012.

[17] R. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[18] Y. Kerr, "Soil moisture from space: Where are we?" *Hydrogeol. Journal*, vol. 15, no. 1, pp. 117–120, 2007.

[19] Y. Kerr, P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M.-J. Escorihuela, J. Font, N. Reul, C. Gruhier, S. Juglea, M. Drinkwater, A. Hahne, M. Martin-Neira, and S. Mecklenburg, "The SMOS mission: New tool for monitoring key elements of the global water cycle," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 666–687, 2010.

[20] Y. Kerr, P. Waldteufel, J.-P. Wigneron, J. Martinuzzi, J. Font, and M. Berger, "Soil moisture retrieval from space: The soil moisture and ocean salinity (SMOS) mission," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 8, pp. 1729–1735, 2001.

[21] S. Kim, J. van Zyl, K. McDonald, and E. Njoku, "Monitoring surface soil moisture and freeze-thaw state with the high-resolution radar of the Soil Moisture Active/Passive (SMAP) mission," *Proceedings of IEEE International Radar Conference, Washington, DC*, pp. 735–739, 2010.

[22] G. Kirchgassner and J. Wolters, *Introduction to modern time series analysis*, 1st ed. Springer, 2007, 282 p.

[23] P. Lee, *Bayesian statistics: An introduction.* Wiley, 2012, 488 p.

[24] J. Lemmetyinen, "Microwave radiometry of snow covered terrain and calibration of an interferometric radiometer," Ph.D. dissertation, Aalto University, Department of Radio Science and Engineering, 2012.

[25] W. Lin, I. Corbella, F. Torres, N. Duffo, and M. Martín-Neira, "Correction of spatial errors in SMOS brightness temperature images," *2011 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3752–3755, 2011.

[26] P. Maybeck, *Stochastic models, estimation, and control Volume 1.* Elsevier Science, 1979, 445 p.

[27] K. McMullan, M. Brown, M. Martín-Neira, W. Rits, S. Ekholm, J. Marti, and J. Lemanczyk, "SMOS: The payload," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 3, pp. 594–605, 2008.

[28] S. Mecklenburg, M. Drusch, Y. Kerr, J. Font, M. Martín-Neira, S. Delwart, G. Buenadicha, N. Reul, E. Daganzo-Eusebio, R. Oliva, and R. Crapolicchio, "ESA's soil moisture and ocean salinity mission: Mission performance and operations," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1354–1366, 2012.

[29] E. Middelberg and U. Bach, "High resolution radio astronomy using very long baseline interferometry," *Rep. Prog. Phys.*, vol. 71, no. 6, 2008, 32 p.

[30] National Radio Astronomy Observatory web site, available (accessed on 21.10.2014): https://science.nrao.edu/.

[31] M. Nemoto, T. Hirota, and Y. Iwata, "Application of the Extended Force-Restore model to estimating soil-frost depth in the Tokachi district of Hokkaido, Japan," *Journal of Agricultural Meteorology*, vol. 64, no. 3, pp. 177–183, 2008.

[32] G.-Y. Niu and Z.-L. Yang, "Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale," *Journal of Hydrometeorology*, vol. 7, no. 5, pp. 937–952, 2006.

[33] A. Papoulis, *Probability, random variables, and stochastic processes*, 3rd ed. McGraw-Hill College, 1991, 666 p.

[34] Y. Pawitan, *In all likelihood: Statistical modelling and inference using likelihood*, 1st ed. Oxford University Press, 2001, 543 p.

[35] J. Ramesh, *Machine vision*, 1st ed. McGraw-Hill College, 1995, 549 p.

[36] K. Rautiainen, J. Kainulainen, T. Auer, J. Pihlflyckt, J. Kettunen, and M. Hallikainen, "Helsinki University of Technology L-band airborne synthetic

aperture radiometer," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 3, pp. 717–726, 2008.

[37] K. Rautiainen, J. Lemmetyinen, J. Pulliainen, J. Vehviläinen, M. Drusch, A. Kontu, J. Kainulainen, and J. Seppänen, "L-band radiometer observations of soil processes in boreal and subarctic environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1483–1497, 2012.

[38] K. Rautiainen, J. Lemmetyinen, M. Schwank, A. Kontu, C. Ménard, C. Mätzler, M. Drusch, A. Wiesmann, J. Ikonen, and J. Pulliainen, "Detection of soil freezing from L-band passive microwave observations," *Remote Sensing of Environment*, vol. 147, pp. 206–218, 2014.

[39] S. Rise, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 24, pp. 46–156, 1945.

[40] D. Riseborough, "Estimating active layer and talik thickness from temperature data: implications from modeling results," *Proceedings on Ninth International Conference on Permafrost*, vol. 2, pp. 1487–1492, 2008.

[41] S. Särkkä, *Bayesian filtering and smoothing.* Cambridge University Press, 2013, 252 p.

[42] M. Schwank, K. Rautiainen, C. Mätzler, M. Stähli, J. Lemmetyinen, J. Pulliainen, J. Vehviläinen, A. Kontu, J. Ikonen, C. B. Ménard, M. Drusch, A. Wiesmann, and U. Wegmüller, "Model for microwave emission of a snow-covered ground with focus on L-band," *Remote Sensing of Environment*, vol. 154, pp. 180–191, 2014.

[43] M. Schwank, A. Wiesmann, C. Werner, C. Mätzler, D. Weber, A. Murk, I. Völksch, and U. Wegmüller, "ELBARA II, an L-band radiometer system for soil moisture research," *Sensors*, vol. 10, no. 1, pp. 584–612, 2010.

[44] G. Seber and A. Lee, *Linear regression analysis*, 2nd ed. Wiley, 2012, 584 p.

[45] SMOS Barcelona Expert Centre web site, available (accessed on 30.6.2014): http://www.smos-bec.icm.csic.es/smos_satellite_views.

[46] H. Sorenson, "Least-squares estimation: from Gauss to Kalman," *IEEE Spectrum*, vol. 7, no. 7, pp. 63–68, 1970.

[47] G. Topp, J. Davis, and A. Annan, "Electromagnetic determination of soil water content: Measurement in coaxial transmission lines," *Water Resources Research*, vol. 16, no. 3, pp. 574–582, 1980.

[48] F. Ulaby, R. Moore, and A. Fung, *Microwave remote sensing: Active and passive. Volume I: Fundamentals and Radiometry.* Artech House Publishers, 1986, 608 p.

[49] F. Ulaby, R. Moore, and A. Fung, *Microwave remote sensing: Active and passive. Volume II: Radar remote sensing and surface scattering and emission theory.* Artech House Publishers, 1986, 608 p.

[50] V. Zaitsev and A. Polyanin, *Handbook of exact solutions for ordinary differential equations*, 2nd ed. Taylor and Francis, 2002, 815 p.

APPENDIX A: CANNY EDGE DETECTION RESULTS

The seasonal changes in the brightness temperature data measured by the ELBARA-II instrument and the SMOS satellite were searched using the Canny edge detector algorithm described in Algorithm 1. In the high latitudes of Sodankylä the soil is frozen continuously for several months, the freezing and thawing periods are distinct and repetitive freeze-thaw cycles are rare. Therefore, a reasonable assumption would be to locate only two edges per year in the brightness temperature data, one which occurs in the autumn and another one in the spring, corresponding to freezing and thawing of the soil respectively. For this reason, the scale at which we are searching for edges is quite coarse and only one threshold, denoted with T, will be used. Several combinations for the value pairs of σ^2 and T were tested. The results obtained for horizontal and vertical polarizations are shown below. The results from ELBARA-II measurements are given first and the results from the SMOS data follow.

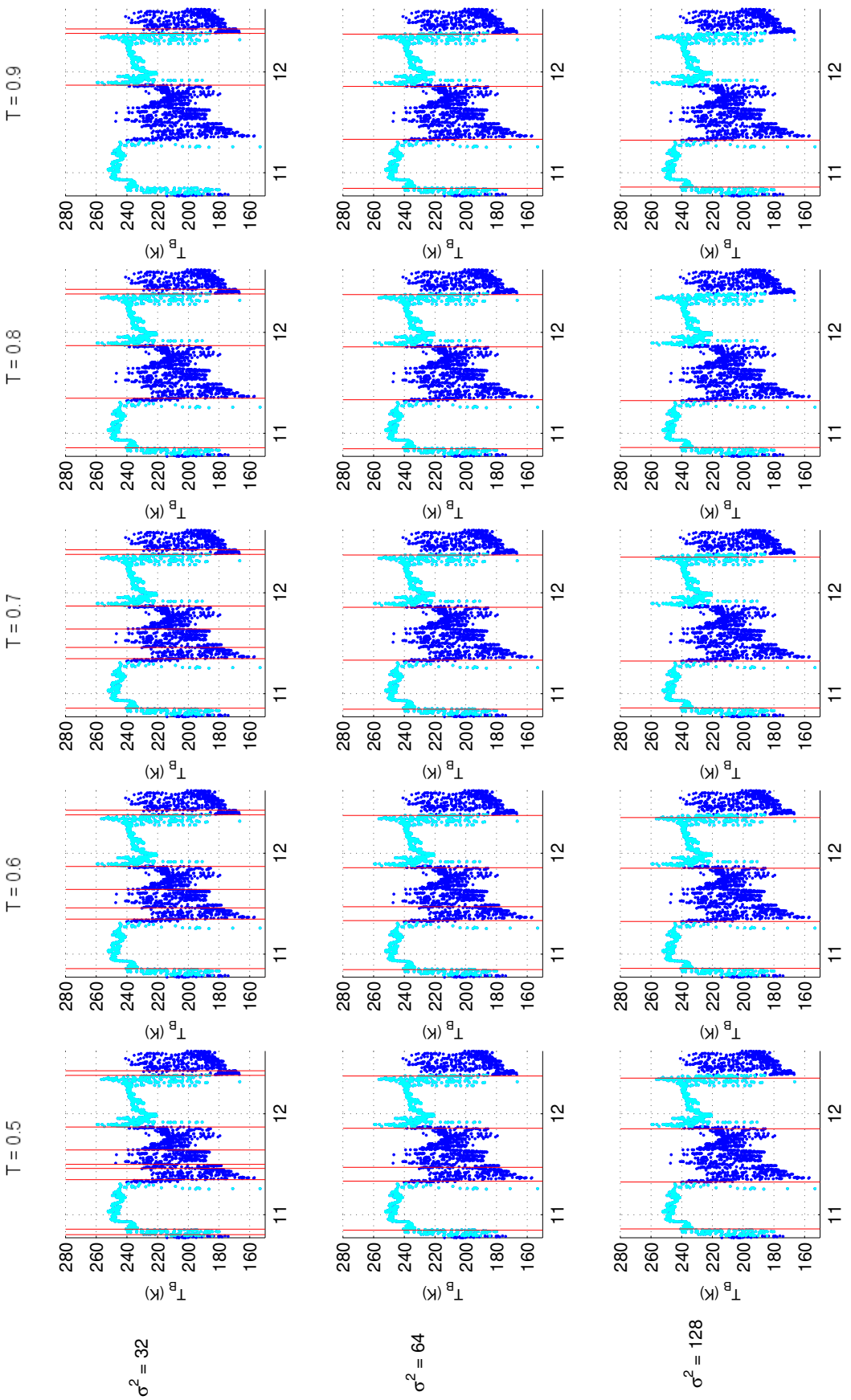

Figure A.1 *Edges found from the ELBARA-II brightness temperatures at horizontal polarization. The x-axis ticks mark the beginning of the year. The winter period is plotted with cyan dots and the edge with a vertical red line.*

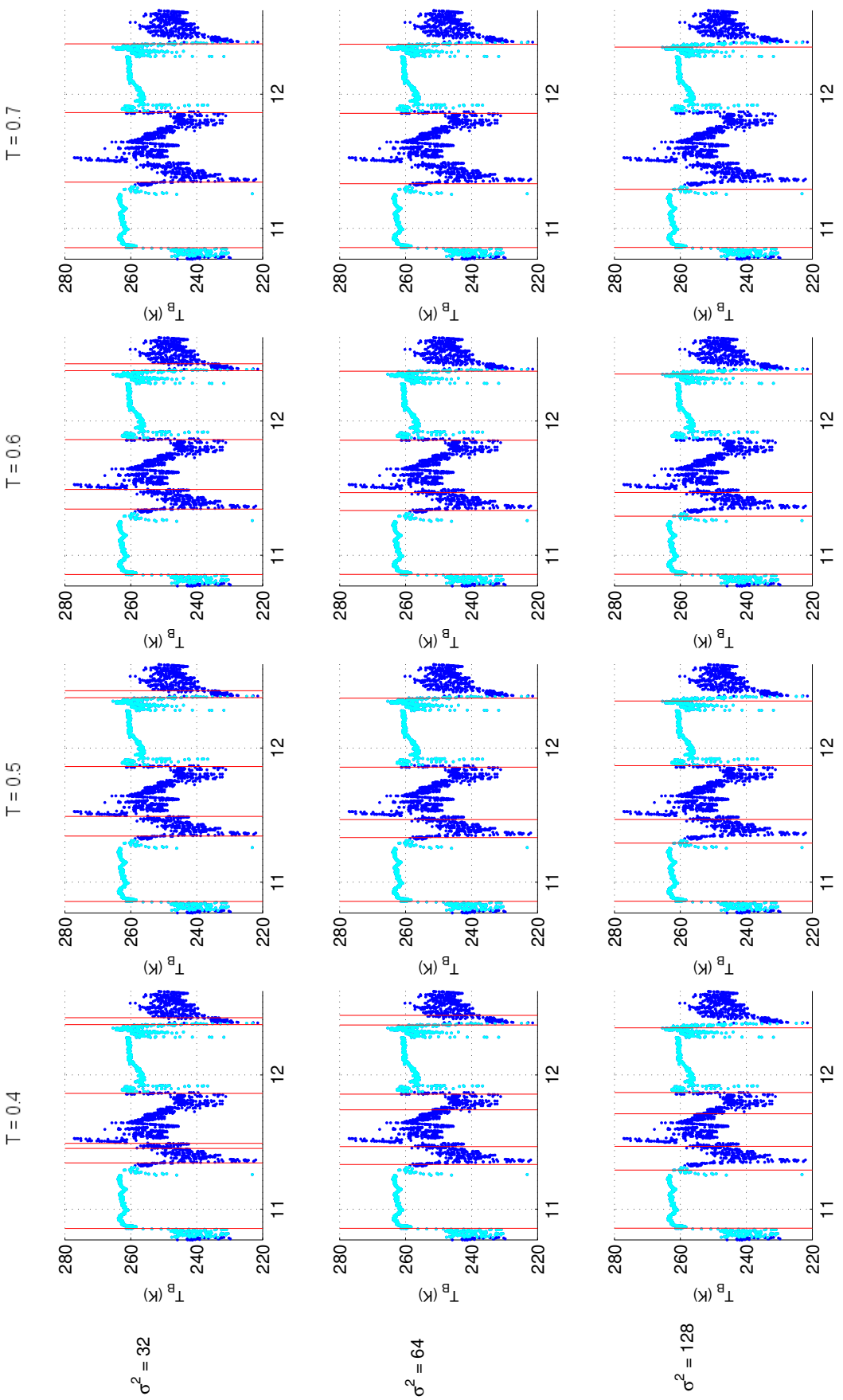

Figure A.2 *Edges found from the ELBARA-II brightness temperatures at vertical polarization. The x-axis ticks mark the beginning of the year. The winter period is plotted with cyan dots and the edge with a vertical red line.*

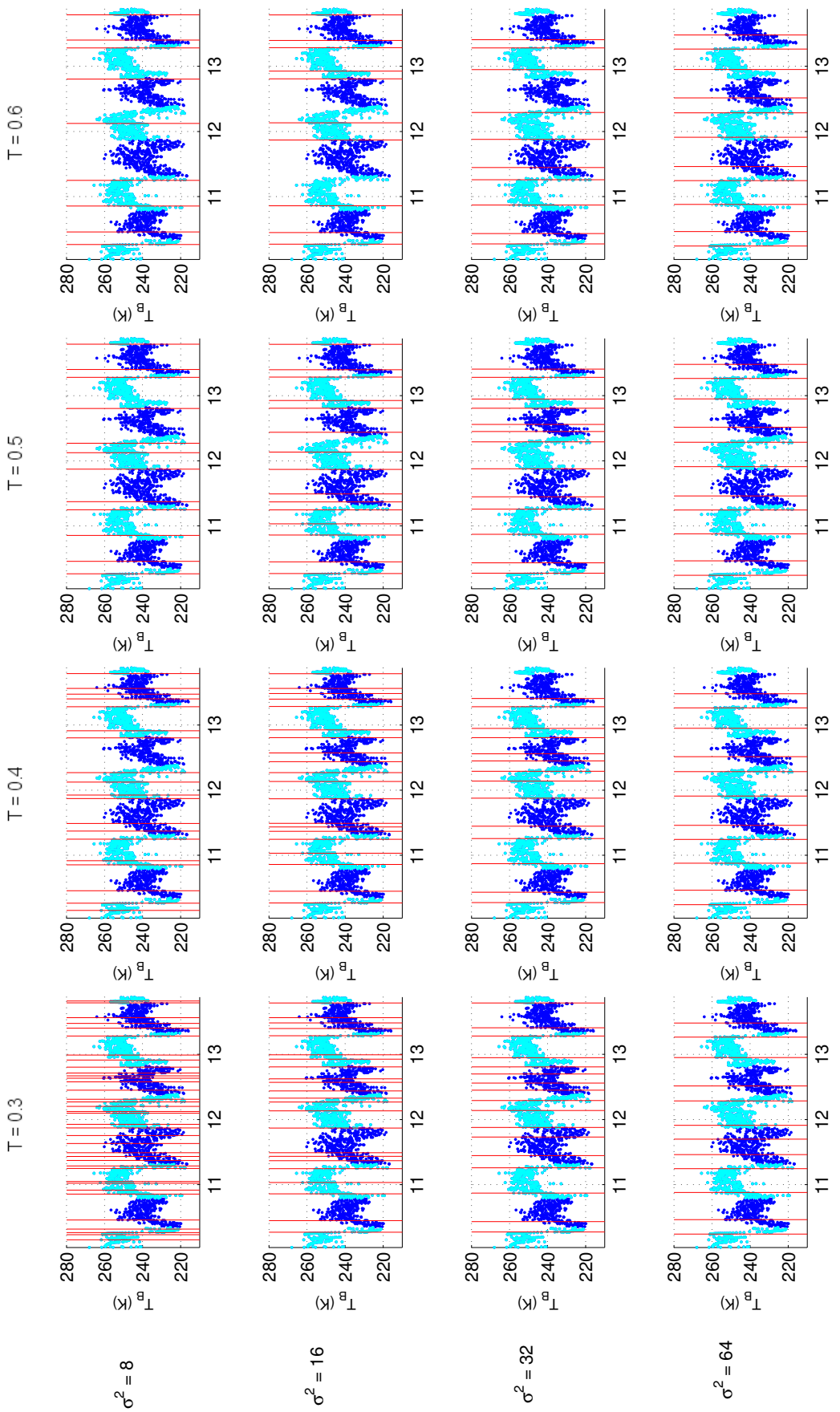

Figure A.3 *Edges found from the SMOS brightness temperature data at horizontal polarization. The x-axis ticks mark the beginning of the year. The winter period is plotted with cyan dots and the edge with a vertical red line.*

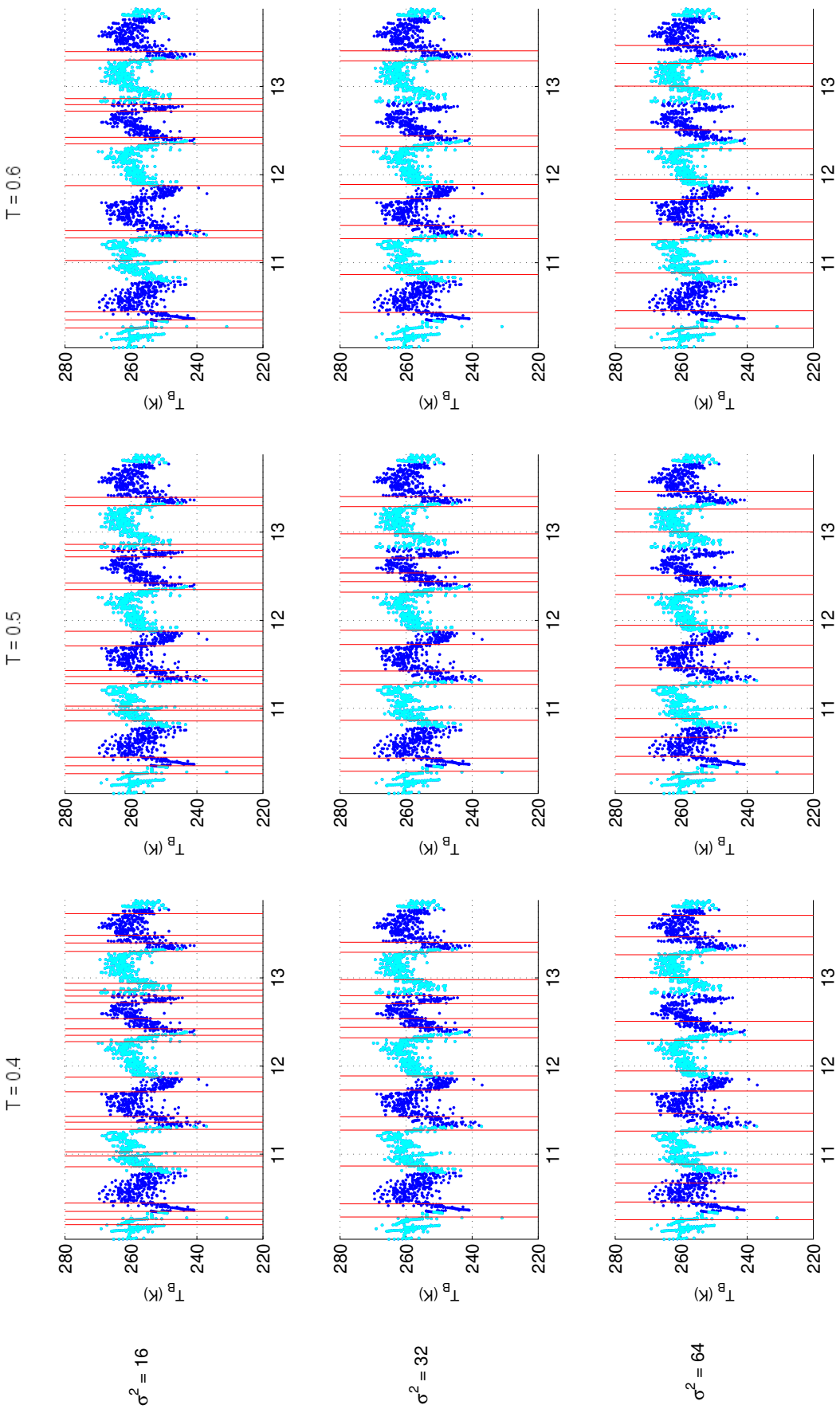

Figure A.4 *Edges found from the SMOS brightness temperature data at vertical polarization. The x-axis ticks mark the beginning of the year. The winter period is plotted with cyan dots and the edge with a vertical red line.*

APPENDIX B: REGRESSION ANALYSIS RESULTS

Three different approaches were tested to gain estimates on the frost and thaw depths from the brightness temperature data measured by the ELBARA-II radiometer. In the first, an algorithm introduced in [38] and discussed in Section 4.2 was applied to the moving average filtered brightness temperature data. In the second, the same procedures were applied to Kalman filtered brightness temperature data. The third and last approch was to formulate two polynomial models, one for the freezing period and another one for the thawing period to see if these simpler models would be adequate to describe the freezing and thawing periods' processes. The frost and thaw depth estimates computed using these three models were further divided into three soil state categories. The computed soil state estimates based on the data filtered with moving average filter and the data filtered with Kalman filter are shown in Figure B.1. The soil state estimates computed for the polynomial models are shown in Figure B.2.

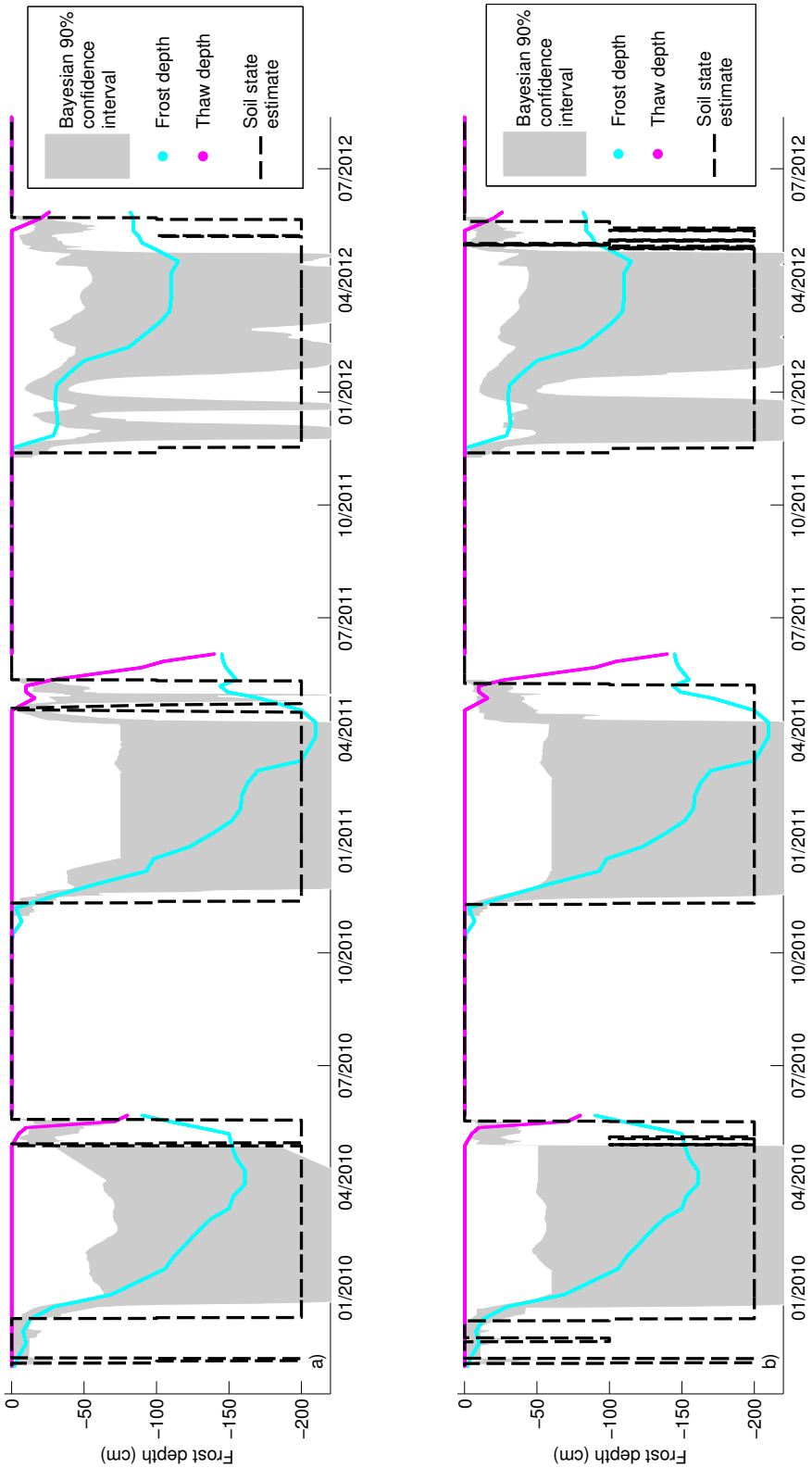

Figure B.1 *a) The soil state estimate based on FF_{rel} computed from the data filtered with moving average filter shown as the dashed line. The shaded area represents the computed 90% Bayesian confidence interval for the estimated frost depth. The measured frost and thaw depths are shown in cyan and magenta respectively. b) The soil state estimate based on FF_{rel} computed from the data filtered with Kalman filter shown as the dashed line. The shaded area represents the computed 90% Bayesian confidence interval for the estimated frost depth. The measured frost and thaw depths are shown in cyan and magenta respectively.*

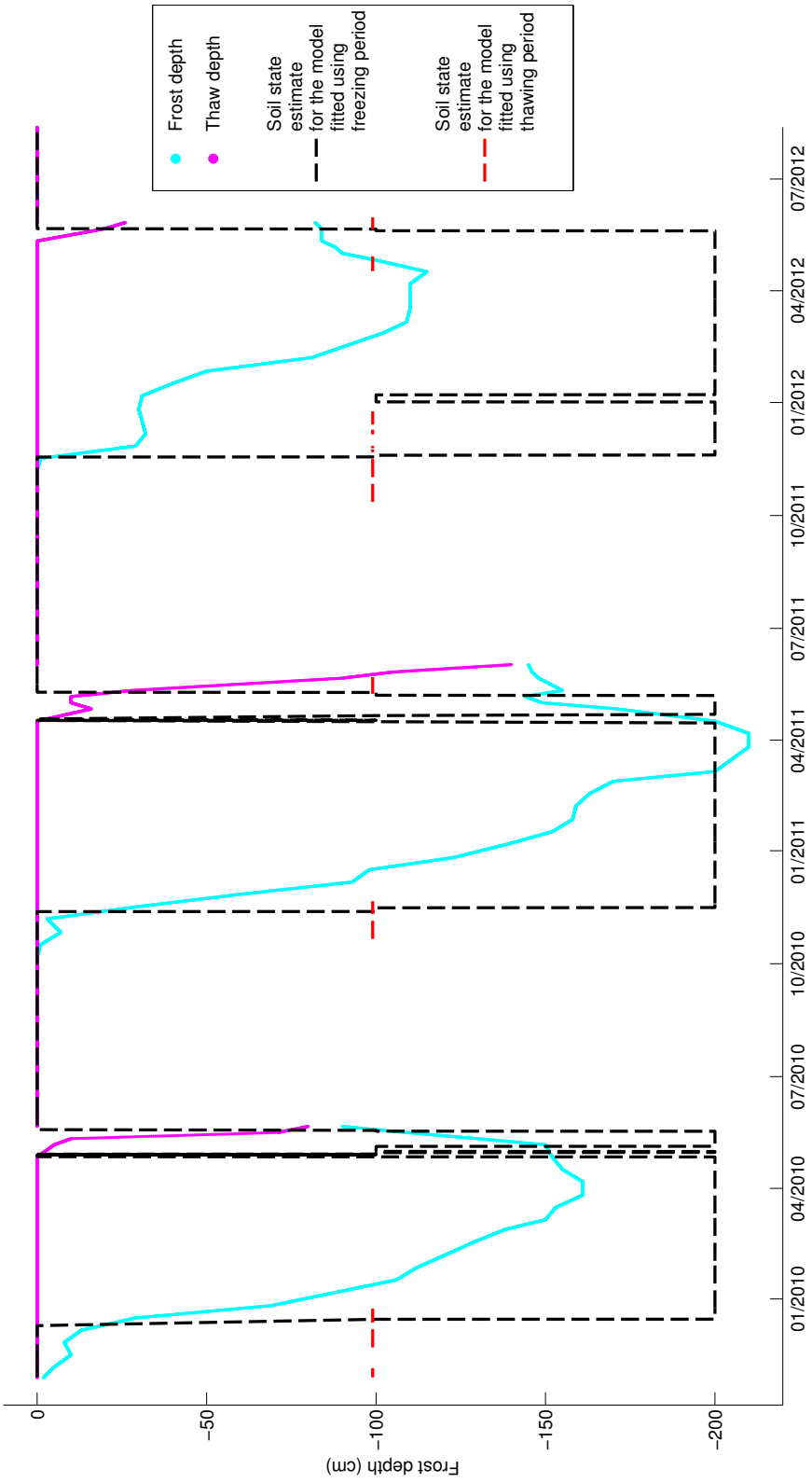

Figure B.2 *The soil state estimates based on the polynomial models. For the model fitted using the thawing periods only the detected freezing and thawing periods are shown. These are plotted at value -99 so that the two soil state estimates would not be on top of each other. Still in the spring 2010 the soil state estimate based on the model fitted using the thawing periods is hardly visible. The measured frost and thaw depths are shown in cyan and magenta respectively.*