The state of the earth: estimating physical parameters from noisy and incomplete earth observation data

Arp, L.R.

1

INTRODUCTION

Earth system parameters are all around us. These physical parameters are scientific variables, whose values describe the Earth at a certain time and place. For example, the Leiden office in which this thesis is being written on a spring afternoon could be described by the following parameters: a temperature of 22.4 degrees Celsius, an air humidity of about 50%, a leaf area index of 0 (sadly; this parameter describes the concentration of the leaves of vegetation) and a $PM10$ fine dust concentration of about 17 $\mu g/m^3$.

Of course, this set of parameters P can differ per use case. Whereas atmospheric scientists may be interested in parameters such as CO_2, NO_2, $PM10$ and $PM2.5$ concentrations in the atmosphere [1, 2], oceanographers may focus on sea surface salinity, pH and ocean wind speed [3] and environmental biologists may care more about chlorophyll $a+b$ content, soil moisture and leaf area index (LAI) [4]. The unifying factor differentiating these physical parameters from conventional variables is that they are measurable quantities representing the state of a physical system governed by physical laws. In our case, this physical system is the Earth system: a highly complex physical system consisting of multiple sub-systems including meteorological processes, ocean dynamics and ecosystems, with possible interactions between these sub-systems.

Although the relevance of these parameters to scientific applications is clear, one pertinent question would be how we can know the correct values of these parameters for our situation in the first place, like the parameter *configuration* $\boldsymbol{\theta} = [22.4, 0.5, 0.0, 17]$ described above for an office at Leiden University. In fact, many of the parameters we may be interested in cannot be directly measured. Others may be directly measurable, but doing so may be costly in terms of human

labour, financial investment, or damage to the very system we are trying to measure (for example, cutting down trees to measure their properties in a laboratory). As an alternative to this *in-situ* approach to data collection, we could instead opt to *estimate* or *infer* these parameters indirectly from data sources that are both measurable and plentiful. When working with Earth system parameters, arguably the most appealing of such data sources is spaceborne Earth observation data: information on the planet obtained by satellites orbiting the Earth, such as the Sentinel satellites from the European Space Agency (ESA) and the Landsat satellites by the American National Aeronautics and Space Administration (NASA).

Earth observation (EO) data can be obtained in various forms; we provide a brief overview hereof in Section 2.1. However, the main focus of our methods is on spaceborne EO data in the form of satellite data. EO satellites continuously orbit the Earth, and transmit their measurements back for further analysis. However, the quantities being measured are generally electromagnetic waves, measured in the form of optical light spectra denoting the intensity of light at the wavelengths of b spectral bands. Many of the physical parameters describing the Earth impact how sunlight gets reflected back into space. This allows us to use light spectra, measured by EO satellites, as feature data to estimate the values of physical parameters on Earth, such as the concentration of gases in the atmosphere or chlorophyll in plant matter. In some fields using EO data, this task is referred to as *parameter retrieval*; throughout this thesis, we will use the standard machine learning terminology of the *estimation* of the physical parameters as target variables, unless stated otherwise.

Definition 1.1 (physical parameter estimation)**.** The estimation of the values $\boldsymbol{\theta}$ of a set of physical parameters (target variables) P that describe the state of the Earth, based on an observed feature vector $\mathbf{x}$ of Earth observation data. The estimated values are denoted as $\hat{\boldsymbol{\theta}}$.

Parameter estimation, therefore, is an inference problem where the configuration of true values for P, $\boldsymbol{\theta}$, must be inferred from the features $\mathbf{x}$: $\hat{\boldsymbol{\theta}} = f(\mathbf{x})$. The mapping function $f : \mathcal{X} \rightarrow D_P$, where $\mathbf{x} \in \mathcal{X}$ (observation domain; in our case, these are light spectra with b bands in $\mathbb{R}^b$) and $\boldsymbol{\theta} \in D_P$ (domain for the parameters P), is generally unknown in advance. For example, in astronomy, we know how large celestial bodies might bend beams of light, because they are derived from well-studied fundamental causal properties of gravity. However, when merely observing a beam of light, it can be challenging to infer whether there was a celestial body that changed its trajectory, let alone which specific object, and the nature of its impact. As will become apparent in Section 1.1, there are challenges associated with parameter estimation using EO data that are specific to this problem setting.

Although we will maintain generality in most of our methods, we will have a particular focus on the estimation of *biophysical parameters* (ecological variables describing the state of an ecosystem) throughout this work as the main use case, specifically those related to vegetation. These parameters, such as leaf area index, chlorophyll content and water content, play a vital role in the health of vegetation and its role in the ecosystem, which, in turn, can strongly affect other Earth systems (e.g., the impact of vegetation and algae on climate through photosynthesis).

1.1. MOTIVATION

In addition to the typical challenges associated with inference problems, the specifics of the problem setting for parameter estimation result in the following two core technical challenges:

1. Many sources of data are inconsistent, resulting in spatial and temporal gaps in the dataset (missing data) and potential biases. For example, sensor networks can only measure a target variable at a few specific locations, and cloud cover can introduce substantial gaps in satellite imagery.

2. The inference problem could be ill-posed due to insufficiently informative or noisy input data, meaning that multiple solutions could be equally valid for a given input. For example, if two parameters affect light, and therefore the satellite imagery, in the same way (e.g., they both increase near-infrared reflectance), it is unclear which parameter the observed patterns should be attributed to.

These challenges, while differentiating parameter estimation using EO data from conventional inference- and machine learning tasks, such as image classification or house price prediction, are likely to be present in a class of similar inference problems, whose shared properties with EO-based parameter estimation result in the same challenges. Applications in such related problems may benefit from the contributions contained in this work.

Inference problems are likely to contain similar challenges to those presented in this work if they share (a subset of) the following properties: i) the input $\mathbf{x}$ consists of noisy, real-world data, ii) the ground truth data $\boldsymbol{\theta}$ is difficult to measure precisely, iii) there are (non-uniformly distributed) data gaps in the input $\mathbf{x}$ and/or the ground truth data $\boldsymbol{\theta}$, such as missing groups of pixels in images, and iv) the data generating mechanism in the real world consists of causal, physical relationships from the parameters $\boldsymbol{\theta}$ to the outcomes $\mathbf{x}$, but in the inference problem, $\mathbf{x}$ forms the observations while $\boldsymbol{\theta}$ are the target parameters (inverse modeling). Intuitively, fields such as astronomy [5, 6] and medical imaging [7, 8, 9] are likely to

share many of these characteristics with the Earth science settings considered in this work.

1.1.1. CHALLENGE 1: DATA INCONSISTENCY

Data inconsistency can affect both the ground truth parameter data $\boldsymbol{\theta}$ and the EO feature data $\mathbf{x}$.

The target variables P, whose values $\boldsymbol{\theta}$ we are aiming to estimate, are inherently spatio-temporal. A parameter describing the Earth does so for a specific location at a certain time. However, the datasets we have available to train or evaluate predictive models may not have measurements for the full region of interest. For example, a sensor network (see Section 2.1.1) takes measurements at various locations in a region, but the number of locations is necessarily limited. Similarly, sensor defects, transmission errors, inhospitable terrain and other unpredictable factors can prevent measurements from being taken for every possible location in a study area. This results in both systematic and random spatial gaps in the ground truth data $\boldsymbol{\theta}$. Similarly, not all measurements in a dataset may have been taken at the same time, or the temporal resolution of measurements could be low (e.g., due to the cost of measurement), resulting in temporal gaps in the data. These spatio-temporal data gaps can make it challenging to train and evaluate predictive models that, for example, return a full image where every pixel represents a prediction for the corresponding geographical location at the same time.

To improve data consistency and enable the training and evaluation of prediction models, we need a method to interpolate the spatio-temporal gaps in the data. The interpolated data should reflect local spatial structure, such as similar neighbours or abrupt changes, as well as global spatial structure, such that spatial interactions can take place over any distance. This led to research question 1 (RQ1), which aims at developing a method that satisfies these requirements:

RQ1 *How can we effectively interpolate spatial data such that both local and global spatial properties are retained?*

In contrast to the ground truth data described above, the EO data $\mathbf{x}$ is usually already a full satellite image covering a geographical area, while the regular orbit ensures consistent revisits (although certain orbits may prioritise frequent revisits over particular geographical areas). Nonetheless, this type of data is still strongly affected by spatio-temporal data gaps. Unlike the sensor networks described above, which often have a consistent pattern to the data gaps, missing EO data is generally caused by unpredictable factors. These factors include invalid pixels (e.g., due to solar glint), sensor faults (e.g., Landsat ETM+ SLC-off data [10]), transmission errors and, most prominently, cloud cover (for details, see Section

2.1.3). The spatio-temporal data gaps caused by these factors are often not uniformly distributed. For example, cloud cover can be affected by the season (more clouds in winter than in summer) and geographical location (more clouds in tropical regions than in deserts). Therefore, this data inconsistency forms a major obstacle for the training of predictive models, and can introduce biases in the trained models. Furthermore, without the feature data $\mathbf{x}$ available at a specific time and place of interest, direct parameter estimation for this spatio-temporal point will not be possible with a typical prediction model mapping features $\mathbf{x}$ to parameters $\boldsymbol{\theta}$.

To address these limitations, we need a method to interpolate missing data in our input EO data $\mathbf{x}$, thereby greatly improving the consistency of our estimations. We took a particular interest in cloud cover, as one of the most prominent causes of missing EO data, but the method should generalise to other types of missing data. Unlike general spatial interpolation problem settings, most satellite-based EO data comes in the form of images. This results in an image processing gap filling task, where the spatial relationships between pixels can have a far greater intensity and variability than in general spatial interpolation settings, while requiring greater precision. Moreover, this spatial structure may change over time, to which the method should be robust. It should be easy to apply the method to any type of EO image data, without the need to re-train a model for every possible type of satellite, and the resulting reconstructed images should be of high quality, such that they can be used in downstream parameter estimation tasks. This leads to the following research question:

RQ2 *How can we effectively and easily interpolate unpredictable, spatially clustered missing data in Earth observation imagery?*

1.1.2. CHALLENGE 2: NOISE AND ILL-POSEDNESS

When performing parameter estimation as $\hat{\boldsymbol{\theta}} = f(\mathbf{x})$, regardless of the method used to approximate f, we are implicitly assuming that the function f is a one-to-one mapping from $\mathcal{X}$ to D_P. Any uncertainty on $\hat{\boldsymbol{\theta}}$ would then be caused solely by flaws in the approximation of f (e.g., the errors of a machine learning model). However, it may be the case that there is not enough information contained in $\mathbf{x}$ to uniquely estimate the correct $\boldsymbol{\theta}$. As a result, f could be a one-to-many mapping, where many different $\hat{\boldsymbol{\theta}}$ could be the appropriate solution for the same $\mathbf{x}$. For example, if the observations were generated as $\mathbf{x} = \boldsymbol{\theta}^2$, and we observed $\mathbf{x} = 100$, there would be two possible answers: $\boldsymbol{\theta} = 10$ or $\boldsymbol{\theta} = -10$. This is one of the ways in which a parameter estimation problem can be *ill-posed* (see Section 2.2.2 for details).

Alternatively, noise on the observations $\mathbf{x}$ can result in ill-posedness. As explained in Section 2.1.3, the EO input data is necessarily noisy (for example, due

to atmospheric interference in the optical data). Consequently, the parameter estimation result $\hat{\boldsymbol{\theta}}$ we obtain as a best fit for this noisy data may not be the true $\boldsymbol{\theta}$ representing the state of the Earth. Instead, any configuration that could have been the parameter estimation result, if the random noise on the EO observations had been different, could possibly be the true solution. Without knowing the noise on the observations, it will not be possible to judge with certainty which is the true solution, making a commitment to any particular solution inappropriate and resulting in ill-posedness.

These examples illustrate how multiple factors could result in ill-posedness for parameter estimation. These factors may reduce the reliability of parameter estimation results, but their impact is not well understood, complicating the development of strategies to alleviate them. Therefore, a thorough analysis of the impact of these factors on the solution reliability is necessary. From this analysis, it should be clear what the factors are, and how they can be alleviated. This is the scope of the following research question:

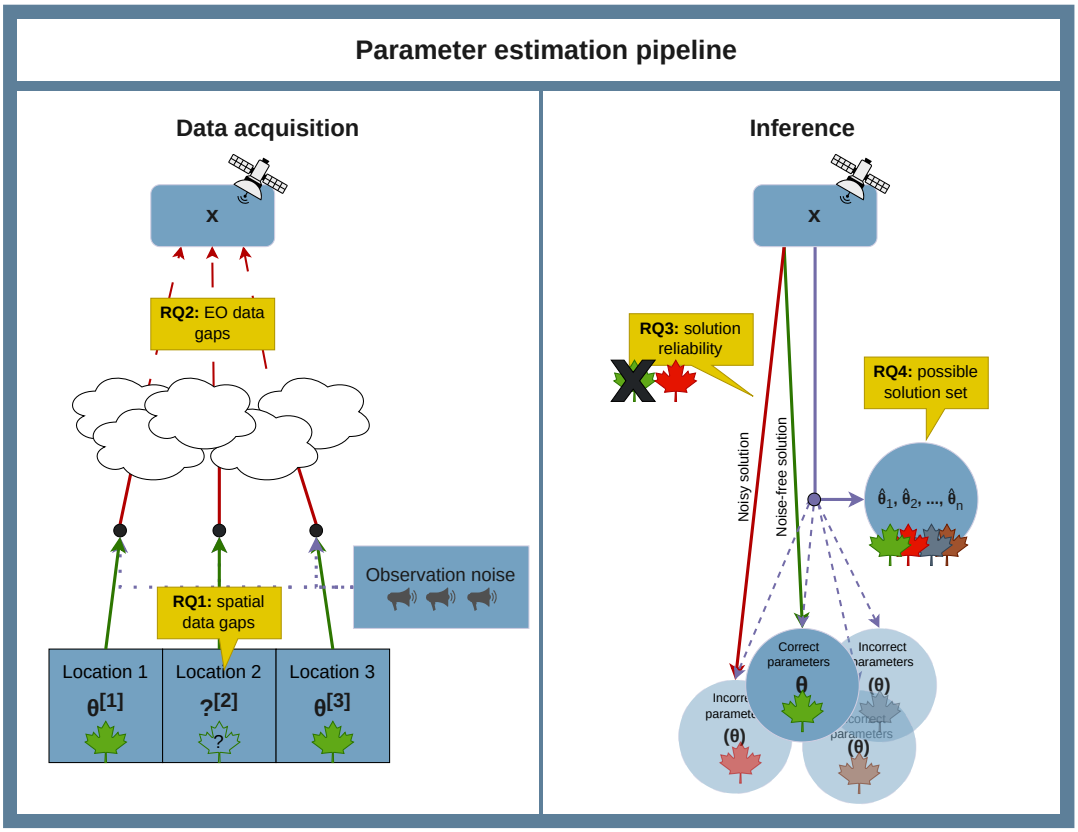
RQ3 *What makes parameter estimation an ill-posed problem, and which factors affect the reliability of parameter estimation results?*

When a problem is ill-posed, finding appropriate solutions can be extremely challenging, but this is not necessarily the case. Even problems known to be ill-posed could be solved at a satisfactory level of accuracy, depending on the problem setting. For example, if there are only a few possible solutions $\hat{\boldsymbol{\theta}}$, all at a negligible distance from each other in the parameter space D_P, the ill-posedness is unlikely to cause major issues for the parameter estimation task. However, if there are many solutions spread around D_P, each of which could be an appropriate solution for the observed EO data $\mathbf{x}$, performing a reliable estimation may be infeasible based on the information available.

Therefore, we need a method to automatically extract all the possible solutions to noisy inference problems. The set of solutions should be accurately approximated, and should include solutions that could have been the parameter estimation result if the noise had been different. This would then allow us to, for example, establish how severely ill-posed a parameter estimation problem instance is. Our final research question, therefore, is as follows:

RQ4 *How can we automatically extract the set of possible solutions to a noisy inference problem?*

An overview of the research questions, and the challenges they address, is shown in Figure 1.1.



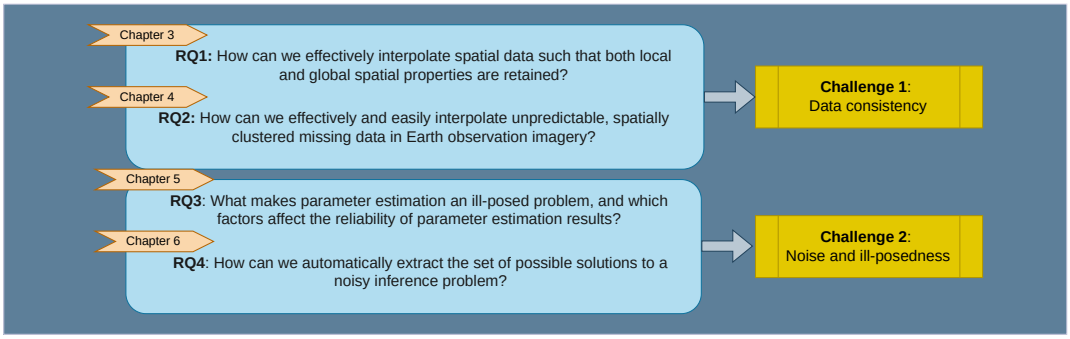


Figure 1.1: Overview of the research questions, the chapters of this work addressing them, which part of the parameter estimation pipeline they address, and the challenges they correspond to. In RQ1, the parameter values $\boldsymbol{\theta}$ are known for some locations (1 and 3), but are unknown for another (2), necessitating a spatial interpolation method for the target values. In RQ2, the satellite measurements of the light reflected by the parameters are blocked by factors such as cloud cover, necessitating a missing data interpolation method for satellite data. In RQ3, random noise causes the spectrum to change shape, resulting in the wrong parameter configuration being returned as a solution, and necessitating an investigation into noise and ill-posedness. In RQ4, there are many possible spectra that random noise could create, resulting in multiple potential solutions and necessitating a method that finds all these potential solutions.

1

1.2. CONTRIBUTIONS

The work covered in this thesis will present the following key contributions to the artificial intelligence and remote sensing communities:

- To address RQ1, we created a spatial interpolation method, called VPint, that provides a novel perspective on spatial interpolation problems and exploits covariates with minimal assumptions. The key idea of our proposed method is its system-oriented perspective, enabling global spatial interactions through a chain of local spatial relationships. We performed extensive empirical experiments on synthetic data, as well as real-world datasets on GDP levels and COVID-19 incidence rates, and compared the performance of VPint to that of representative baseline methods including Kriging methods, spatial autoregressive models, and convolutional neural networks. We found VPint to perform better than competing methods on average in terms of mean absolute error (MAE), root mean squared error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). This contribution is contained in Chapter 3, and has given rise to the following journal article:

  > Laurens Arp, Mitra Baratchi, and Holger H. Hoos. (2022). *VPint: value propagation-based spatial interpolation.* Data Mining and Knowledge Discovery, 36:1647–1678. Springer.

- To address RQ2, we adapted the VPint algorithm to VPint2, which applies the concepts of VPint to the image processing-like problem of cloud removal in Earth Observation data. We also created a cloud removal benchmark dataset, SEN2-MSI-T, enabling a validation of our proposed method and supporting the development of future cloud removal methods. VPint2 can be easily applied without prior training and requires only a single reference image of the same sensor. The modifications and extensions we made allowed the algorithm to be successfully applied to cloud removal and EO data, which introduce particular complications distinct from general spatial interpolation problems, due to variable spatial autocorrelation structures that could be dynamic over time. We compared our proposed method to baseline methods including mosaicking, automated machine learning, neighbourhood similar pixel interpolation and previously published cloud removal deep learning models, and found it to perform significantly better in terms of mean absolute error (MAE), mean absolute percentage error (MAPE), structural similarity index (SSIM) and downstream task performance in 17 out of 20 conditions. This contribution is contained in Chapter 4, and has given rise to the following journal article:

> Laurens Arp, Holger H. Hoos, Peter van Bodegom, Alistair Francis, James Wheeler, Dean van Laar, and Mitra Baratchi. (2024). *Training-free thick cloud removal for Sentinel-2 imagery using value propagation interpolation.* ISPRS Journal of Photogrammetry and Remote Sensing, 216:168–184. Elsevier.

- To address RQ3, we carried out a thorough empirical study aimed at characterising the ill-posedness of vegetation parameter estimation through radiative transfer model (RTM) inversion (for details on RTM inversion, see Section 2.2.1). We systematically tested whether the formal properties of ill-posedness (see Section 2.2.2) are met when performing RTM inversion. Next, we performed experiments aimed at gaining insight into the relationship between noise on the observed EO data and ill-posedness. Finally, we tested the mechanisms through which common strategies to reduce ill-posedness are effective. Our empirical experiments found that RTM inversion met all the requirements of a well-posed problem, but that noise on the EO data resulted in ill-posedness for the parameter estimation problem. Based on this knowledge, we recommend future work to focus on data-centric contributions, such as improving the quality of the EO data. This contribution is contained in Chapter 5, and has given rise to the following journal article:

  > Laurens Arp, Peter van Bodegom, Holger H. Hoos, and Mitra Baratchi. (2026). *Characterising the Ill-posedness of PROSAIL Inversion for Biophysical Parameter Retrieval.* European Journal of Remote Sensing, 59(1). Taylor and Francis.

- To address RQ4, we formalise the concept of ϵ-manifolds, and propose eMMI, a method to efficiently approximate the ϵ-manifold. Based on our findings on ill-posedness, we found that noise on the observations of an inference problem can cause the inference problem to become ill-posed. An ϵ-manifold is a set of all possible solutions to an inference problem, whose suitability is characterised by a loss function value threshold ϵ. Therefore, approximating the ϵ-manifold enables us to automatically extract the set of all possible solutions to a noisy inference problem. Our proposed approximation method, eMMI, efficiently performs a search around the point prediction based on diversity optimisation, and trains a classifier on the sampled points. We validated our proposed method on simulation models including RTMs, simulation-based inference simulators and machine learning tasks, and compared against statistical methods such as Gaussian processes, Bayesian neural networks and approximate Bayesian computation. Our empirical experiments found that ϵ-manifolds are significantly better than statistical uncertainty quan-

tification at containing the true solution, and eMMI approximated the ϵ-manifold significantly better than the baseline methods. This contribution is contained in Chapter 6, and has given rise to the following journal article:

> Laurens Arp, Peter van Bodegom, Nguyen Dang, Holger H. Hoos, Alistair Francis, and Mitra Baratchi. (2025). *Inference from Noisy Observations through Model Inversion: Constructing ϵ-Manifolds of Potentially Valid Solutions.* Under review.

All contributions in this thesis are accompanied by the code required to use the methods developed, as well as the code required to reproduce our results. Other resources, such as datasets created for the projects, can either be downloaded directly, or can be reproduced using the provided scripts.

1.3. ORGANISATION OF THIS DISSERTATION

The rest of the dissertation is structured as follows. In Chapter 2, we provide the necessary background information to allow readers to put our work in context. The following four chapters (Chapters 3 – 6) cover our key contributions, following the structure of the research questions (RQs) and contributions described in Sections 1.1 and 1.2. We conclude the dissertation in Chapter 7 with a discussion on the work contained therein, and recommendations for future research directions. Since the interdisciplinary nature of this work virtually guarantees that readers will be unfamiliar with parts of the subject matter, we provide a glossary of key terms at the end of this dissertation.