Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Master Thesis

Improving the Availability of Contextual Data with Machine Learning-Based Interpolation

Ian Maurice Buck

ian.buck@studium.uni-hamburg.de
Study Program Information Systems
Matr.-Nr. 6911467

First Reviewer: Prof. Dr. Janick Edinger
Second Reviewer: Philipp Kisters

Abgabe: 08.2023

A distributed system is one where the failure of some
computer I've never heard of can keep me from getting my work done.
– Leslie Lamport

Abstract

Many science-based applications need continuous or gridded input data in order to work properly. This paper investigates how single data-points can be combined to create a continuous data layer, in which missing data points are interpolated. An example for such an application would be the prediction of temperatures coming from single sensors and weather-stations, that can be combined to detect Urban Heat Island (UHI)s. UHIs are weather phenomena that get ampflified among other things by the ongoing densification of urban areas to create more living space, typically accompanied by the removal of green areas that can help with the dissipation of heat. These heat islands, in which the temperature is significantly higher than in surrounding areas, pose a threat to the health of the urban population, especially to the elderly, children and people with existing health conditions. Traditionally, UHIs are detected using Land Surface Temperatures (LST) captured by satellites, that usually have the downside of low spatial and temporal density. This paper proposes an alternative approach by creating a machine-learning model that is able to interpolate missing data between data-points coming from citizen-owned sensor networks that are combined with mobile sensors, which can be attached to rental bikes, buses or e-scooters, to gain temporary insights into otherwise unobserved areas. The model combines data streams of sensor readings with historic data and creates a fine-granular continuous data-layer, in this case for temperature, which allows for an accurate localization of UHIs.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LST . . . . Land Surface Temperatures

UHI . . . . Urban Heat Island

1 Introduction

In 2023 56% of the human population already lives in urban areas with the number projected to continiously increase to 68% by 2050 [1]. Combined with the ongoing climate change and urban densification, cities are facing many new challenges. With the removal of vegetation in favor of living space and the sealing of surfaces with heat-absorbing materials such as asphalt or concrete for streets and highways [GRGTDW20], rising temperatures lead to new phenomena that pose risks for the urban citizens. A recent phenomenon is the appearance of so called urban heat-islands (UHI). A heat-island is a local occurrence of significantly higher temperatures than surrounding areas that pose a health risk, especially for the elderly, children or citizen with prior health-issues [MBG15].
In order to detect UHIs, Land Surface Temperature (LST) is commonly used. While allowing for a cheap analysis of large areas without the need of ground weather-stations, this approach comes with certain downsides, such as low temporal and spatial resolution and restrictions such as only being able to measure temperatures when no clouds interfere with the microwaces send from the measuring satellite [ZPL15]. This spatial and temporal resolution, of f.e. spatial resolution of 0.01° longitude and 0.01° latitude (equal to roughly 1.11km by 1.11km) and temporal resolution of monthly average surface temperature as offered by LST data provided by the European Space Agency (ESA) Climate Office's data set [GVP], is not enough to effectively analyze the urban microclimate. Another candidate that comes to mind are weather stations. They usually provide hourly, for current values sometimes even 10 min interval readings of temperature, rain and wind, but don't offer the necessary spatial resolution. Lastly, there is the possibility of deploying sensor networks to closely monitor the climate of the city, but this approach can be quite cost intensive for a large amount of sensors over a long time period [CMY+15]. An alternative that is less costly would be to instead rely on citizen-owned sensor networks from the existing Smart Home and Internet of Things (IoT) infrastructure, like Sensor.Community [2] and Netatmo [3], which offer a temporal resolution of 5 min for temperature and wind, hourly for rain, while also having a comparably high spatial resolution. This approach has the desired temporal resolution and has been shown to work well in [MFG+17], but there might be areas, such as industrial zones, where citizens are not able or not allowed to install their personal sensors. In order to also gain insights in such previously unobserved areas, we propose the usage of mobile sensors

[1] https://ourworldindata.org/urbanization#by-2050-more-than-two-thirds-of-the-world-will-live-in-urban-areas
[2] https://deutschland.maps.sensor.community/
[3] https://weathermap.netatmo.com/

that could be installed on buses, bikes or e-scooters to gain temporary snapshots and improve the spatial resolution even further. As research realted activities commonly rely on continuous or gridded data fields, there needs to be a way to convert these single data points from the different sensors into a continuous data-layer.
In this paper, we propose a solution to this problem by training a machine learning regression model, that allows for the interpolation of missing data-points. Based on sensor readings, from the sensor networks and mobile sensors, of commonly collected weather information, such as temperature, humidity, rain, pressure, wind, and possibly other variables such as vegetation indexes [AR20], the model then creates a continuous data-layer that allows for a holistic view of the observed variable, in this case temperature.

To do (??)

1.1 Objective

current situation: abundance of data (data quality unknown), but different data sources at different places with different formats, making it hard to work with different sources at the same time - smart city example -> heat island detection - sensor networks (statinary + mobile) - LST satellite data (lack of spatiotemporal resolution, not suited for micro-climate) - vegetation indexes (in geoinformation systems/portals) - etc. - currently these sources are used independently from each other, but how can they be integrated? - hybrid approaches have shown that combining different approaches (smart city stationary + moving sensors) to give better prediction quality than singular approach (reasons: unobserved areas) - statistical models offer not enough flexibility/are to cumbersom to work with (and probabilities are not known) - ML is a good fit to analyse patterns and rules - but effort to retrain models for each application expensive - how can ML be used to integrate the different types of data? - how can data availibility of contexttual data be improved by leveraging ML approaches for different topics? - for the heat map detection, what you need is a fine-granular temperature map, that allows for outlier detection - either train your own model on a huge dataset - or access a temperature map -> this will be discussed, how to use ML for this and improve data availibilty by using interpolation - additonal topics to briefly discuss (mainly apply by implementing ML models) - what work needs to be done before using the data in ML (preprocessing, transformation, outlier detection etc.) - one way of preparing data is to interpolate missing data to create continous gridded features (focus of this work) - how ML can we improve the interpolation quality of features? -> turn sparse features into denser versions with interpolation, interpolation based on other features present

Methodology

In order to validate these things, we need high quality data-sets with many different features, but there are not many available, even tho for certain things (atmosphere etc.)

there are OpenData sets available. Goal is: implement a temperature map based on ML that interpolates missing data based on historical and current data. Compare prediction quality of different type of models (DL, random forests etc.) based on cross-validation (simulate missing data by leaving data points out on purpose) -> experiment with different intervals, densities etc. -> could also compare to Krigin (as a typical geostatistical approach for temperature interpolation) Subgoal: In order to achive this, we also need high quality data-sets to train and validate the model -> source them from OpenData platforms

1.2 Structure of this work

Research methods:
- literature research as foundation
- heat island detection - smart cities -> sensor networks vs LST - what type of other data is available? geoinformation, vegetatition, for micro-climate: shades of bigger buildings?
- - prototyping
- implement pipeline to pre-process different types of data
- feature extraction
- implement ML model
- deploy ML model

- create/search for fine granular data sets
- add new contextual data to exitsing data sets
- discuss different types of data (gridded vs continous vs data points) and methods for each
- train ML models with different methods (deep learning, random forests etc.) and different features enabled

- cross validation of results
- discuss validation techniques and indicators (RSME, MSE)

The rest of the thesis is structured as follows. Chapter 2 begins with an analysis of related work, where important literature is discussed, that forms the foundation of this research. In chapter 3, the focus lies on describing the service architecture, that shows how a machine learning model can be deployed in different contexts to improve data availability. Which machine learning approaches can be used to interpolate missing data and how they differ from each other is discussed in chapter 4. In Chapter **??**, the different ML approaches are compared and cross validated with each other based on the different model that are trained on the obtained data-sets. Finally, chapter **??** discusses the findings of this thesis and gives an outlook into future work and research directions.

2 Related Work

In the following chapter we lay the foundation for the research conducated in this thesis. We start off with an introduction of the topic of urban heat island (UHI) detection, which is one of the motivating factors behind this work. UHI research is part of the discipline of modern climatology and relies on many different sub-disciplines like meteorology, thermal dynamics, geology and many more. After discussing traditional ways to detect UHIs, such as using remote satellite data, we discuss newer approaches, especially in the context of smart cities, that focus on collecting air temperatures in urban environments. Measuring data in the urban environment comes with many challenges [Oke06], that can lead to missing or wrong observations for given places. In order to improve data availibility in the urban climate context we compare traditional interpolation techniques based on (geo-)statistics with more recent machine-learning based approaches, that could be used in this highly complex urban setting. This interpolated data could be used to improve interpolation or prediction models and give additional insights into the influences of urban environments, f.e. in the context of UHIs and their detection and analysis. Finally, in order to compare traditional geostatistical models with ML-based models, we need comprehensive data-sets of urban climate data. An overview of available data sources is given in section 2.3 together with an introduction to the OpenData movement.

2.1 Urban Heat Islands (UHI)

UHIs have been the center of a lot of attention for quite some time in the scientific community. As early as 1833, with the research of Luke Howard in London which observed higher temperatures inside London than in surrounding areas [How33], UHIs have seen a steadily increase in scientific contributions. The term *Urban Heat Island* was first introduced in the 1940s [BP47]. The recording and investigation of UHIs has seen mayor steps since the begin of modern climatology, also known as the Sundborg's era beginning with Sundborg's 1951 classic heat island study of Uppsala [Sun51]. UHIs occur in many cities around the globe [PPC+12] in different climatic zones, during different times of day and in different intensities.

Shared UHI Challenges

Some shared challenges are: 1. Define what *urban* means in the context of UHIs [SO09]. The term urban is widely used to identify areas that are more densly populated than the surrounding rural areas. Having this distinction between urban and rural [Low77] helped researchers to better define the UHI magnitude (cite), but this simple distinction also lead to problems [Ste11]. The problem lies in the fact that there is no clear border between urban and rural areas, but a fluent transition. Especially for larger metropolitan areas, like Tokyo, the urban area could span 10s to 100s of kilometers, making the collection of reference rural temperatures hard. The reference rural temperature has a direct influence on the UHI magnitude, which is 'the most widely recognized indicator of city climate modification in the encironmental sciences' [SO09]. As a solution, different classification into local climate zones were proposed [SO12, SO09], that classify areas based on surface roughness, building densities, building heights etc. 2. Measuring the influence of other local urban or meterological phenomena on the temperatures collected. The urban climate is extremly complex, due to many different influences, such as antropological energy,... (cite, todo find the influence factors). Additionally, the urban climate is also influenced by surrounding regional/meso-scale climate phenomena such as storms, valleys, mountains, large waterbodies, costlines and more (cite).

2.1.1 UHI Classification

UHIs can be classified in many different ways. Typically, there is a horizontal classification, defining the superficial extension of the UHI from micro-, to local- to meso-scale, and a vertical classification, defining in which vertical layer of the urban area the heat island is observed. To better understand these scales and the anatomy of the planitory/urban boundary layer, figures 2.1 to 2.3 show a detailed view of the meso-, local- and micro-scale of the urban climate, as illustrated by Oke 2006 [Oke06].

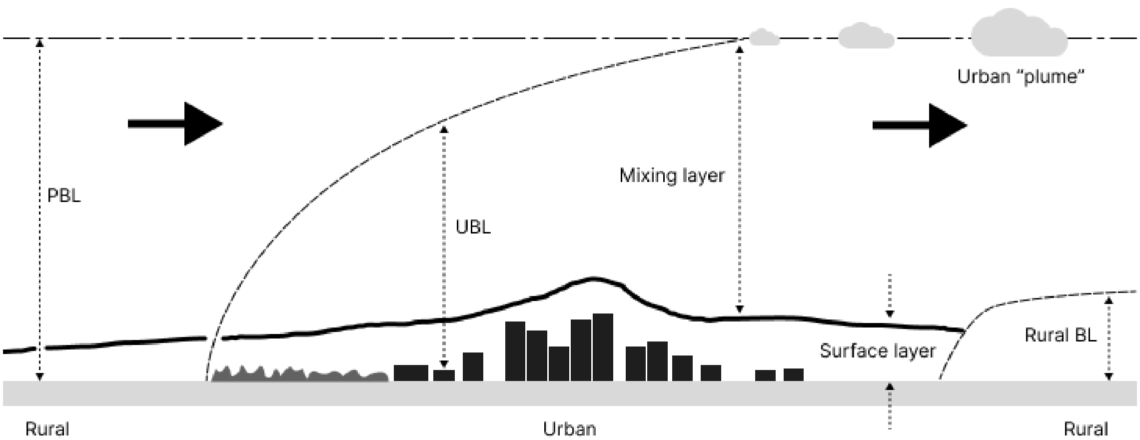

Figure 2.1: Mesoscale view of the urban climate, redrawn from [Oke06]

The mesoscale, as depicted in fig. 2.1, spans the whole urban environment of a city, typically tens of kilometres. There are several boundary layers, that comprise different scales. The planetary boundary layer (PBL) [Wyn85] is the lowest layer of the Earth's atmosphere and spans from the surface to a height of several hundred meters up to several kilometers. It is characterised by the turbulent mixing of air, forming wind currents, that are mayorly influenced by the underlying surface.

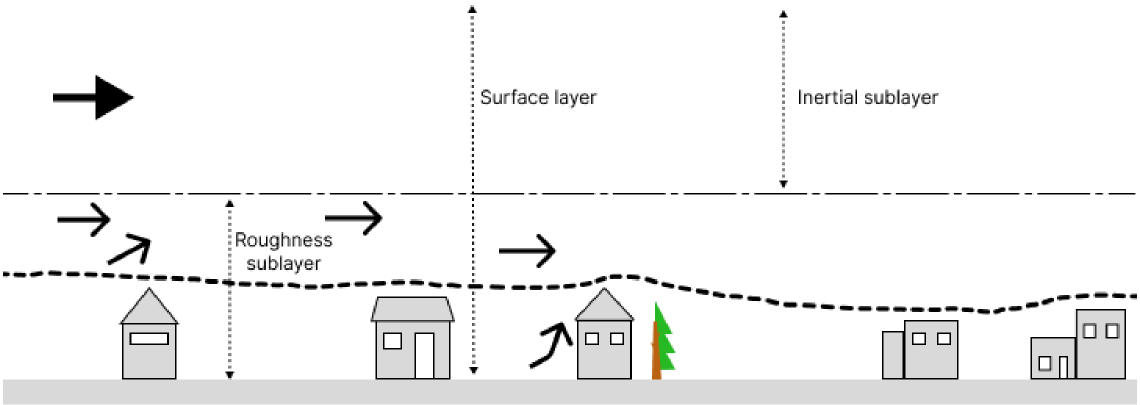

Figure 2.2: Localscale view of the urban climate, redrawn from [Oke06], (Todo finish)

The localscale is situated closer to the surface and contains landscape features such as topography, but does not yet include microscale effects. At this layer, the underlying microclimatic effects in form of fluxes mix together to form a more average and representative view of the source area, typically at the scale of one to several kilometers. This layer is monitored by weather stations that are located at/or sligthly above the canopy height. Todo: more infos

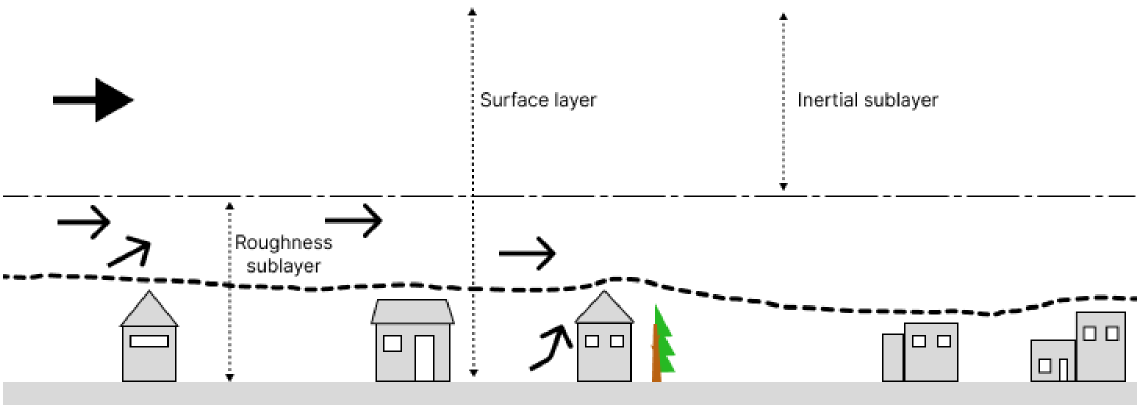

Figure 2.3: Microscale view of the urban climate, redrawn from [Oke06], (Todo)

The microscale deals with the characteristics of each individual surface area. On of the main measurements on this layer is the surface temperature, which is primarily influenced by solar radiation. It is important to note, that surface and air temperatures are correlated, but this correlation varies greatly based on other surrounding influences such as wind velocity or humidity [SB92]. todo: more

Vertically UHIs can be devided into three mayor types [Oke76, OMCV17], namely Boundary Layer Heat Island (BUHI), Canopy Urban Heat Island (CUHI) and Surface Urban Heat Islands (SUHI), that correspond with the boundary layer they can be monitored/measured in. todo: more

2.1.2 Canopy Urban Heat Island (CUHI)

The canopy UHI is measured in the canopy boundary layer several meters above ground slightly below or on the average roof layer of the surrounding buildings, as seen in fig. 2.3. The primary measurement in the canopy is air temperature, which is used to measure the urban heat island intensity (UHII) [Oke73], the most commonly used way of describing the heat island maginitude [KB21].
Since the beginning of modern climatology, mayor progress has been made in this research field, but methodologies and scientific rigor in CUHI research still seems to be lacking, as discussed by Stewart in 2011 [Ste11]. Stewart found, that over 54% of CUHI research was lacking proper methodologies or had other shortcommings such as a lack of site desciptions, where sensors were placed, or the disregard of non-urban factors such as local weather phenomena. In response, progress has been made in recent years by improving methodologies and ensuring correct measurements of climate-related data and study design and execution through various guidelines [Oke06], especially in urban settings, that require special care due to the huge amount of possible influences on local recording sites.

Todo: more

2.1.3 Surface Urban Heat Island (SUHI)

The surface temperature is measured directly at the surface of an object and is the main indicator of the surface urban heat island (SUHI). Surface temperature, in contrast to air temperature, typically is measured via remote sensing technologies like LST via satellites. Well-known satellites include MODIS, Landsat ... (todo list all), that all carry different types of instruments and sensors, that are able to take various measurements. Through the use of satellites, the spatial coverage is great, but raster sizes usually range from one kilometer to a hundred, so the spatial resolution is not that great compared to denser sensor networks. Additionally, these satellites are not geostionary to cover a wide range and therefore only take measurements during a handfull of times a day (todo: number of fly overs with example). Another downside is that satellites in many cases need clear-sky condidtions to measure surface temperature at ground level, as their sensors are not able to penetrate the cloud surface. As a solution, some technologies such as LIDAR

(todo check if true) offer the approximation of the underlying surface temperature below clouds by measuring the out-going radiation of the surface. todo: List of measurement types (microwaves etc.) with disadvantages
Surface and air temperatures can vary greatly, therefore a SUHI does not necessarily also imply a CUHI.

2.1.4 UHI Detection in the Context of Smart Cities

Smart Cities are 'urban areas that exploit operational data [...] to optimze the operation of city services' [HEH+10], by collecting near-real-time data from physical and virtual sensors, integrating those data sources into an enterprise computing platform and performing complex analytics on them. With current urbanization trends (60% ofd ppl living in cities 2060) and the ongoing global and urban warming (cite), research into smart cities has gotten a lot of attention recently (cite). On of the driving factors behind smart cities is next to progress in digitisation of cities the availibility of cheap smart sensors (cite), that enable a good spatiotemporal surveillance of factors in a smart city.
- goals and pillars - architecture - challenges
  - classification of UHI detection into the smart city framework
  In the context of UHI detection,
  - testbeds, sensor networks (citizen owned)
  - cross over to pollution detection

2.2 Interpolation of Missing Data

In the context of urban environments, there are many measurements such as surface temperature, which are measured by sensors that have certain weaknesses. In the case of LST data collection, clouds play a mayor part and prevent surface temperature to be measured. In many LST data-sets (ref), measurements for cloud areas are simply defined as having no value. As many applications and algorithms need continous input data to work as expected, in this work we take a look on how interpolation, especially with the help of ML, can help solve this problem. In the context of LST data, this could mean interpolating the missing data either based on surrounding data (cite) or by integrating many different features such as air temperature, humidity, heat flux etc. in a ML model (cite).
- moving sensors for better spatial coverage into unobserved areas

Interpolation vs. Extrapolation

In this work, we focus on interpolation of missing data. Interpolation is the process of calculating/guessing missing values between given values. This could be a missing value in a time-series or a missing cell in a data grid (cite). In contrast, extrapolation is the prediction of values based on previous values, like predicting values based on historic time-series data, es in the case of weather forecasting. Due to time constraints, we only focus on the interpolation part, tho extrapolation could also play an important role in smart cities by predicting UHIs and warning citizens about potential future heat-related risks.

2.2.1 Regression Analysis in Statistics

Fundamentaly, regression analysis has its roots in mathematics, more specifically in the field of statistics. todo: more

- foundation of other research fields, based in statistics/mathematics

- linear regression (least-squares) - multiple regression models - hierachical regression

- special cases - piecewise linear regression - inverse prediction - weighted least squares - logistic regression - poisson regression

2.2.2 Interpolation in Geostatistics

- spatiotemporal (kriging) - time series prediction vs interpolation of missing data - based on GIS - pipeline: fit measured data points to grid, interpolate missing squares downside: cannot find local maximas/minimas

2.2.3 Interpolation with Machine Learning Models

- ML regression

2.3 Access to data sets

As part of this thesis, the collection of own data via sensors or other means is not feasible due to time and resource constraints. However, there are many onlie sources and portals/catalogues that give access to a wide variety of data-sets.

2.3.1 OpenData Movement

- portals - official: - EU: https://data.europa.eu/data/datasets?query=temperature&locale=en (combinas many governmental and local catalogues) / https://data.europa.eu/data/catalogues?locale - USA: https://data.gov/ - UK: https://www.data.gov.uk/ - private: https://www.kaggle.com/

- strategies: - self procurement (test beds -> Helsinki Testbed (climate research meso sclae), UK Birmingham Testbed (climate + smart city),)

2.3.2 Available Data Sources and Portals

2.4 Applications and Research Areas (needed?)

- focus on temperature interpolation/climate research

- climate research - high area coverage, low spatio and temporal resolution (5km by 5km squares) - based on LST data (from satelites) -> not the same as air temperature

- micro-climate research - bad/costly area coverage, high spatio and temporal resolution - Urban heat islands - Pollution (fine dust pollution)

- connection to smart cities - integrating many heterogenous data points - detection climatic anomalys - notifiy/warn residents

- important key words

- current status quo

- important authors and current work

-> identify reseach gap - convert single data points to continous/gridded data - improve density of data to gain insights and imrove visibility -> identify areas with low prediction quality

3 System Architecture

The main goal of this work is to compare different interpolation techniques that are based on traditonal (statistical) and ML-based approaches. ML-based approaches first of all need a lot of data to be trained and validated with, and after deployment need access to relevent real-time data, if near real-time capabilities are desired. In the context of smart city, such ML-based interpolation models could be used to improve data availibility by interpolating missing/unavailable data such as LST readings under cloudy conditions, and be incoperated into a service that other services, like a UHI detection service, could further rely on, without the need to implement interpolation techniques themselves. This could reduce costs to develop such depending services, as they no longer need to deal with missing data themselves while also improving interpolation results with well-trained and designed models.

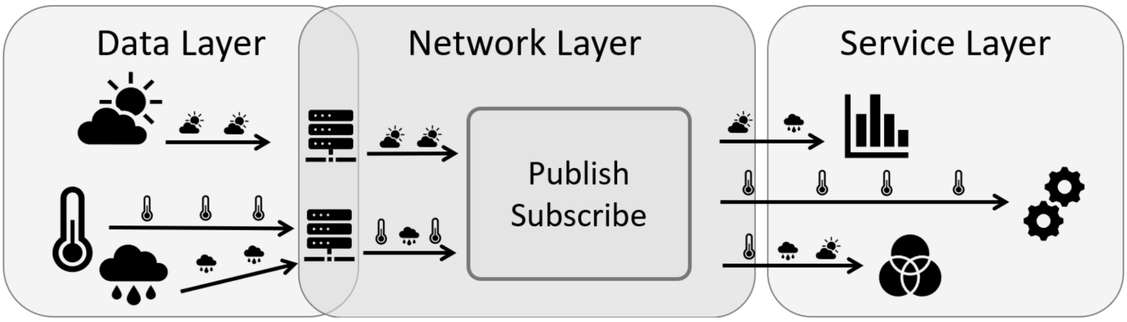

Figure 3.1: In the data layer (left), a wide variety of environmental data is collected with the help of multiple sensors. These are connected to their citizen-owned local base stations, which manage access rights and forward collected data to subscribed services (right) via the decentralized publish-subscribe in the network layer (center).

First, we need to take a look at how such a service fits into existing smart city architectures. The most generalised architecture of a smart city consists of four layer, the sensing layer, transmission layer, data management, and application layer [SKH18]. In this work, we focus on the sensing layer, dealing with topics such as correct sensor placement and underlying footprint, and the application layer, which accesses available data via data management services to provide additional services to the city and its citizens. For the data transmission and data management layers, there already are different technologies and service offerings, that aim at solving the underlying problems, e.g. network bandwidth, network availability, sensor discoverability, handling the massive amounts of data that is already or will be collected in the future, and many more. For the communication and discovery of sensor nodes, one solution could be SkipNet [HDJ+02], an overlay network focused on discoverability while also protecting privacy, with which the data transportation layer could be desinged as a peer-to-peer (P2P) network. Other research focuses on the data accessability and discoverability, by making data accessible for everyone, not only for economic partners in a closed-off system. Examples would be the Smart Networks For Urban Citizen Participation (SANE) initiative [BJK+19], which... Figure 3.1 shows the architecture of the SANE system.

3.1 Sensing Layer

The goal for the sensing layer is to monitor the surrounding environment and capture key data for further analysis and decision making. It consists of many different types of physical and virtual sensors. The first group of sensors are the physical sensors, which are placed directly inside the environment. Wireless sensor networks (WSN) have seen a lot of attention for many different applications such as 'military sensing, physical security, air traffic control, traffic surveillance, video surveillance, industrial and manufacturing automation, distributed robotics, environment monitoring, and building and structures monitoring' [CK03]. The challenges for WSNs primarily depend among other things on the deployment. An ad-hoc WSN has energy and bandwidth contraints due to the usage of batteries as power sources. In Contrast, sensors that are permanetly installed, either stationary or on a moving target, and connected to a constant power source don't have this constraints. This approach could be used for smart cities to reduce waste and guarantee representative measurements via correct sensor placement. In the case of stationary sensor networks though, the initial deployment and following maintenance cost can be substantial [CMY+15].
In recent years, low-cost sensors (LCS) in combination with sensor networks have enabled fine-granular real-time monitoring of urban environments, although the quality of individual low-cost sensors can be questionable [CDS+17]. In general, LCSs can improve data availibility and support analysis, but do not substitute well-calibrated reference instrumentation [LPS18].

3.1.1 Stationary Sensor Types

There are many different types of environmental features that can be measured directly inside an urban area. The types of measurements that can be observed are among others: air temperature, humidity, atmospheric pressure, reactive gaseous air pollutant (CO, NOx, O_3, SO_2), particulate matter (PM), greenhouse gases (CO_2, CH_4), percipitation, solar radiation, wind speed and direction, anthropogenic heat, noise, sky-view factor, heat fluxes and many more. Correlations between these features can vary greatly based on surrounding factors. In order to better understand these correlations, many em-

pirical studies have studied the influence of meteological factors on features such as PM [TMJ10]. Additionally, many fields of statistics have specialised on topics such as statistics in climatology [VSZ02], geostatistics [TYU86] and more.
All sensor readings that are taken by physical sensors are singular data points. Additionally to the type of measurement taken and the actual value observed, physical sensor readings include the physical location of the sensor, e.g. latitude, longitude and altitude, the type of sensor used to take the measurement, and the sampling rate. For air temperature, the sampling rate could be an average temperature measured over five minutes, whereas percipitation might be measured by collecting rain for certain periods of time and then measuring the amount of rain collected. The sensor type is important, as different types of sensors can produce different qualities of measurements, e.g. LCS compared to callibrated reference-grade high cost sensors, and perform better or worse based on the meteological conditions, e.g. worse performance at low temperatures, high humidity etc. Due to the placement directly inside the environment, (near) real-time observation and high temporal resolution are generelly possible, but might be infleunced by network availability etc. The spatial resolution highly depends on the number of sensors deployed and the correct placement of the sensors. The correct placement has a direct influence on the footprint of the sensor [LF14] and the representativness of the measurement taken for the underlying and surrounding area [Oke06].

3.1.2 Remote Sensing

In comparison to stationary sensors that are installed directly in the environment they are observing, remote sensing describes the process of observing a target environment from afar [CW11]. In climatology, remote sensing is used to collect meteological data via satellites, planes or ballons by either capturing image data, that can be used to identify things like cloud and land coverage, by measuring passive radiation, or by actively sending out microwaves or using LIDAR to detect features such as surface temperature, e.g. LST data. Remote sensing comes with its own set of challenges.

-

- sensing layer: (wireless) sensor networks, virtual sensors via APIs/middlewares, sensors as database (bad), privacy

- overview types of sensors, challenges for each - discussion premium vs low-cost sensors - discussion static data incorporation -> NDVI from geoportal via data management layer

- P2,5: highly depends on air flows - humidity: needs to be replaced more often due to pollution - temperature

- modelling of the sensing layer in this thesis: collect actual data (maybe lack of sensor descriptions), simulation (but very complex weather models)

- overview of remote sensing sensor types and downsides (clouds etc.)

- discussion virtual sensors/static content

The *sensing layer* consists of many different data sources. In the context of temperature sensing and prediction, this could include single (inexpensive) sensors such as the popular BMP280 [1], private weather stations such as sold by Netatmo [2] hidden behind an API [3], public weather station data such as from the Deutscher Wetterdienst (DWD) [4] which offer an API and historic weather data, or other geologically relevant data such as zoning plans which, in the case of the city of Hamburg in Germany, can be accessed via an Open-Data platform provided by the State Office for Geoinformation and Surveying Hamburg [5]. In order to gain detailed insights into urban microclimates, we need fine-grained spatial and temporal data. As managing and maintaining such a large sensor network as a single entity can be quite challenging and cost intensive [CMY+15], we rely primarily on crowdsourced sensor data, in this case climate-related, from citizens, that give access to their personal sensors that they f.e. installed at home. This approach has been shown to work well in the densely populated urban area of Berlin, Germany [MFG+17], with the main challenge being data quality assessment due to faulty data from either broken, wrongly configured or wrongly installed sensors. The different data sources provide data streams which are then ingested by our interpolation service. Main challenges are the uncertainty in networks, such as single sensors or APIs not being available due to network interruptions, and the integration of many heterogeneous data sources that can contain data in different formats, time intervals or units of measurement etc.
The *network layer* is responsible for integrating these different data sources in a consistent and reliable way and making them available for other services. The different sensors present in the data layer might have different vendors and programming interfaces, be located behind (vendor specific) APIs or are unreliably accessible due to unstable networks in edge environments. The network layer can be designed as a peer-to-peer (P2P) network based on the SkipNet approach [HDJ+02], that utilizes the lookup efficiency of distributed hash tables and adds support for value-based range queries based on prefixes and attribute-value pairs. Another challenge in the context of the network layer, especially in the context of this paper, is also the integration of mobile sensors, which might not be constantly connected to a network while moving.
The data layer is then exposed via a publish-subscribe architecture [BJK+19] to the *service layer*, that offers subscriptions to and consumption of data streams and houses services such as our temperature interpolation service. These services can also build upon one another. An example for such a dependency could be a UHI detection service that relies on the temperature interpolation service and offers real-time detection of UHIs, which could trigger notifications/warnings for citizens living in the specific area.

- architecture of the smart city (sensing layer, transmission layer, data management, ap-

[1] https://www.bosch-sensortec.com/products/environmental-sensors/pressure-sensors/bmp280/
[2] https://www.netatmo.com/en-gb/weather/weatherstation
[3] https://dev.netatmo.com/apidocumentation/general
[4] https://www.dwd.de/
[5] https://geoportal-hamburg.de/geo-online/

plication layer) [SKH18] - existing smart cities solutions (data marketplaces, data portals etc.)

4 Machine Learning Model Design

The data-layer exposes a variety of single data-points for different points in time. As meteorological research and analysis activities are usually in need of gridded or continuous data [SKP+20], in this paper we design a temperature interpolation service that offers continuous temperature data for a given area. This service is part of the *service-layer* and acts as a building block for further temperature related research and analysis, as temperature is an important variable for research in agronomy, meteorology, hydrology, ecology and many other fields of application.
The core of the service is a deployable regression model, that is capable of interpolating missing data points for a target feature in a defined area, based on surrounding data points, that contain various features that are related to the target feature. In the case of temperature interpolation, the target feature is the temperature and the related features could be temperature, rain, humidity, solar radiation or wind, which can be collected using weather stations and specialized sensors. Other features in the context of temperature prediction could also be geological data such as Normalized Difference Vegetation Index (NDVI) or Modified Normalized Difference Water Index (MNDWI), which according to [AR20] have a strong impact on their estimation model.
In order to find the most appropriate model for this application, we want to compare different promising approaches for interpolating data points such as classical geostatistical methods like Empirical Bayesian Kriging (EBK) or EBK-Regression Prediction (EBKRP) [NAEB23] with machine-learning approaches such as random forest regression [AR20]. The goal is to minimize prediction errors and identifying the features that have the biggest impact on prediction quality, like possibly the density of weather stations [NAEB23].
One focus point of this study will also be the underlying uncertainty and dynamic of the data layer. Because the data layer consists of sensors that are connected to a network, there are bound to be times when the network is unreliable and sensors are temporarily not available. This could happen either for bigger areas if a network provider is having an outage in one of its data centers, or more localized, if the Wi-Fi is unreachable for a single sensor. The network is also run by citizens, so each one can decide to turn off their own sensors, exchange them for new ones or add additional sensors to the network. The prediction model of the temperature map service needs to take this into consideration when interpolating the data.

Because we have a highly dynamic underlying data structure that we need to account for, we think that we could improve the prediction results even further by increasing the

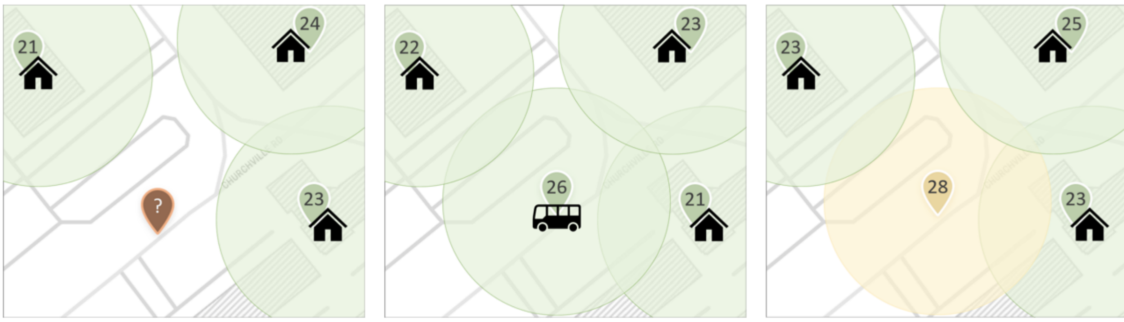

Figure 4.1: Some areas are not covered by stationary sensors (left). Whenever mobile sensors collect data in these areas (middle) this knowledge can be used to train regression model which predicts weather conditions for these unmonitored areas (right).

density of data points and adding readings for previously unobserved areas by adding mobile sensors to the data layer. These sensors could be deployed in an urbanized area by being installed on buses, bikes or e-scooters, which also have the advantage that they usually move close to heat absorbing surfaces such as streets or move through parks which can generally help with heat dissipation. The question here would be, what type of sensors are applicable for such local snapshots and which features could be captured. Temperature can be read easily with a small and inexpensive sensor like the BMP280, whereas collecting rain might be difficult for a moving object.

- compare regression based models

5 Preparation of Datasets

5.1 Automated Data Collection

Build Netatmo api connector, load data

- overview of sources for good datasets for temperature and climate related research

goal: - collect mutiple datasets (many features, fine-granualar spaciotemporal) - enhance datasets with additonal information (soil conditions, zoning plans, vegetation health)

Bibliography

[AR20] ALONSO, Lucille ; RENARD, Florent: A new approach for understanding urban microclimate by integrating complementary predictors at different scales in regression and machine learning models. In: *Remote Sensing* 12 (2020), Nr. 15, S. 2434

[BJK+19] BORNHOLDT, Heiko ; JOST, David ; KISTERS, Philipp ; ROTTLEUTHNER, Michel ; BADE, Dirk ; LAMERSDORF, Winfried ; SCHMIDT, Thomas C. ; FISCHER, Mathias: SANE: Smart networks for urban citizen participation. In: *2019 26th International Conference on Telecommunications (ICT)* IEEE, 2019, S. 496–500

[BP47] BALCHIN, William George V. ; PYE, Norman: A micro-climatological investigation of bath and the surrounding district. In: *Quarterly Journal of the Royal Meteorological Society* 73 (1947), Nr. 317-318, S. 297–323

[CDS+17] CASTELL, Nuria ; DAUGE, Franck R. ; SCHNEIDER, Philipp ; VOGT, Matthias ; LERNER, Uri ; FISHBAIN, Barak ; BRODAY, David ; BARTONOVA, Alena: Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? In: *Environment international* 99 (2017), S. 293–302

[CK03] CHONG, Chee-Yee ; KUMAR, Srikanta P.: Sensor networks: evolution, opportunities, and challenges. In: *Proceedings of the IEEE* 91 (2003), Nr. 8, S. 1247–1256

[CMY+15] CHAPMAN, Lee ; MULLER, Catherine L. ; YOUNG, Duick T. ; WARREN, Elliott L. ; GRIMMOND, C Sue B. ; CAI, Xiao-Ming ; FERRANTI, Emma J.: The Birmingham urban climate laboratory: an open meteorological test bed and challenges of the smart city. In: *Bulletin of the American Meteorological Society* 96 (2015), Nr. 9, S. 1545–1560

[CW11] CAMPBELL, James B. ; WYNNE, Randolph H.: *Introduction to remote sensing*. Guilford Press, 2011

[GRGTDW20] GRÊT-REGAMEY, Adrienne ; GALLEGUILLOS-TORRES, Marcelo ; DISSEGNA, Angela ; WEIBEL, Bettina: How urban densification influences ecosystem services—a comparison between a temperate and a tropical city. In: *Environmental Research Letters* 15 (2020), Nr. 7, S. 075001

[GVP] GHENT, D. ; VEAL, K. ; PERRY, M.: *ESA Land Surface Temperature Climate Change Initiative (LST_cci): Multisensor Infra-Red (IR) Low Earth Orbit (LEO) land surface temperature (LST) time series level 3 supercollated (L3S) global product (1995-2020), version 2.00*. http://dx.doi.org/10.5285/ef8ce37b6af24469a2a4bdc31d3db27d

[HDJ+02] HARVEY, Nicholas J. ; DUNAGAN, John ; JONES, Mike ; SAROIU, Stefan ; THEIMER, Marvin ; WOLMAN, Alec: Skipnet: A scalable overlay network with practical locality properties. (2002)

[HEH+10] HARRISON, Colin ; ECKMAN, Barbara ; HAMILTON, Rick ; HARTSWICK, Perry ; KALAGNANAM, Jayant ; PARASZCZAK, Jurij ; WILLIAMS, Peter: Foundations for smarter cities. In: *IBM Journal of research and development* 54 (2010), Nr. 4, S. 1–16

[How33] HOWARD, Luke: *The climate of London: deduced from meteorological observations made in the metropolis and at various places around it*. Bd. 3. Harvey and Darton, J. and A. Arch, Longman, Hatchard, S. Highley [and] R. Hunter, 1833

[KB21] KIM, Se W. ; BROWN, Robert D.: Urban heat island (UHI) intensity and magnitude estimations: A systematic literature review. In: *Science of the Total Environment* 779 (2021), S. 146389

[LF14] LECLERC, Monique Y. ; FOKEN, Thomas: *Footprints in micrometeorology and ecology*. Bd. 239. Springer, 2014

[Low77] LOWRY, William P.: Empirical estimation of urban effects on climate: a problem analysis. In: *Journal of Applied Meteorology and Climatology* 16 (1977), Nr. 2, S. 129–135

[LPS18] LEWIS, Alastair ; PELTIER, W R. ; SCHNEIDEMESSER, Erika von: Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. (2018)

[MBG15] MARTIN, Philippe ; BAUDOUIN, Yves ; GACHON, Philippe: An alternative method to characterize the surface urban heat island. In: *International journal of biometeorology* 59 (2015), S. 849–861

[MFG+17] MEIER, Fred ; FENNER, Daniel ; GRASSMANN, Tom ; OTTO, Marco ; SCHERER, Dieter: Crowdsourcing air temperature from citizen weather stations for urban climate research. In: *Urban Climate* 19 (2017), S. 170–191

[NAEB23] NJOKU, Elijah A. ; AKPAN, Patrick E. ; EFFIONG, Augustine E. ; BABATUNDE, Isaac O.: The effects of station density in geostatistical prediction of air temperatures in Sweden: A comparison of two interpolation techniques. In: *Resources, Environment and Sustainability* 11 (2023), S. 100092

[Oke73] OKE, Tim R.: City size and the urban heat island. In: *Atmospheric Environment (1967)* 7 (1973), Nr. 8, S. 769–779

[Oke76] OKE, Timothy R.: The distinction between canopy and boundary-layer urban heat islands. In: *Atmosphere* 14 (1976), Nr. 4, S. 268–277

[Oke06] OKE, Timothy R.: Initial guidance to obtain representative meteorological observations at urban sites. (2006), 01, S. 51

[OMCV17] OKE, Timothy R. ; MILLS, Gerald ; CHRISTEN, Andreas ; VOOGT, James A.: *Urban climates*. Cambridge University Press, 2017

[PPC+12] PENG, Shushi ; PIAO, Shilong ; CIAIS, Philippe ; FRIEDLINGSTEIN, Pierre ; OTTLE, Catherine ; BRÉON, François-Marie ; NAN, Huijuan ; ZHOU, Liming ; MYNENI, Ranga B.: Surface urban heat island across 419 global big cities. In: *Environmental science & technology* 46 (2012), Nr. 2, S. 696–703

[SB92] STOLL, Matthew J. ; BRAZEL, Anthony J.: Surface-air temperature relationships in the urban environment of Phoenix, Arizona. In: *Physical Geography* 13 (1992), Nr. 2, S. 160–179

[SKH18] SILVA, Bhagya N. ; KHAN, Murad ; HAN, Kijun: Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. In: *Sustainable cities and society* 38 (2018), S. 697–713

[SKP+20] SEKULIĆ, Aleksandar ; KILIBARDA, Milan ; PROTIĆ, Dragutin ; TADIĆ, Melita P. ; BAJAT, Branislav: Spatio-temporal regression kriging model of mean daily temperature for Croatia. In: *Theoretical and Applied Climatology* 140 (2020), S. 101–114

[SO09] STEWART, Iain ; OKE, TR: Newly developed "thermal climate zones" for defining and measuring urban heat island magnitude in the canopy layer. In: *Eighth Symposium on Urban Environment, Phoenix, AZ*, 2009

[SO12] STEWART, Ian D. ; OKE, Tim R.: Local climate zones for urban temperature studies. In: *Bulletin of the American Meteorological Society* 93 (2012), Nr. 12, S. 1879–1900

[Ste11] STEWART, Iain D.: A systematic review and scientific critique of methodology in modern urban heat island literature. In: *International Journal of Climatology* 31 (2011), Nr. 2, S. 200–217

[Sun51] SUNDBORG, Åke: *Climatological studies in Uppsala: With special regard to the temperature conditions in the urban area.* 1951

[TMJ10] TAI, Amos P. ; MICKLEY, Loretta J. ; JACOB, Daniel J.: Correlations between fine particulate matter (PM2. 5) and meteorological variables in the United States: Implications for the sensitivity of PM2. 5 to climate change. In: *Atmospheric environment* 44 (2010), Nr. 32, S. 3976–3984

[TYU86] TRANGMAR, Bruce B. ; YOST, Russel S. ; UEHARA, Goro: Application of geostatistics to spatial studies of soil properties. In: *Advances in agronomy* 38 (1986), S. 45–94

[VSZ02] VON STORCH, Hans ; ZWIERS, Francis W.: *Statistical analysis in climate research*. Cambridge university press, 2002

[Wyn85] WYNGAARD, John C.: Structure of the planetary boundary layer and implications for its modeling. In: *Journal of Applied Meteorology and Climatology* 24 (1985), Nr. 11, S. 1131–1142

[ZPL15] ZHANG, Xiaoyu ; PANG, Jing ; LI, Lingling: Estimation of land surface temperature under cloudy skies using combined diurnal solar radiation and surface temperature evolution. In: *Remote Sensing* 7 (2015), Nr. 1, S. 905–921

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang XXX selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den ________________ Unterschrift: __________________________

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den ________________ Unterschrift: __________________________