

# Wie sehen Wind Parks aus?

## Das visualisieren globaler Wind Parks

### DIPLOMARBEIT

zur Erlangung des akademischen Grades

### Diplom-Ingenieurin

im Rahmen des Studiums

### Data Science

eingereicht von

**Marina Haller, BSc**

Matrikelnummer 01405809

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ. Prof. Silvia Miksch, PhD, MSc

Mitwirkung: Victor Schetinger, PhD, MSc

Johannes Schreiber, PhD, MSc

Wien, 1. Jänner 2023

---

Marina Haller

---

Silvia Miksch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# What do wind farms look like?

## Visualizing global wind farms

### DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

### Diplom-Ingenieurin

in

### Data Science

by

**Marina Haller, BSc**

Registration Number 01405809

to the Faculty of Informatics

at the TU Wien

Advisor: Univ. Prof. Silvia Miksch, PhD, MSc

Assistance: Victor Schetinger, PhD, MSc

Johannes Schreiber, PhD, MSc

Vienna, 1<sup>st</sup> January, 2023

---

Marina Haller

---

Silvia Miksch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Marina Haller, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. Jänner 2023

---

Marina Haller



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Danksagung

Ich bin mehreren Personen sehr dankbar, die mich während des gesamten Prozesses meiner Masterarbeit unterstützt, angeleitet und inspiriert haben. Zunächst möchte ich mich bei Dr. Victor Schetinger und Dr. Johannes Schreiber für ihre unschätzbaren Beiträge bedanken. Johannes' ursprüngliche Idee für dieses Projekt war die treibende Kraft hinter meiner Forschung, und Dr. Schetingers kontinuierliche Unterstützung waren entscheidend für die Ausrichtung und den Umfang meiner Studie. Ihr Feedback und ihr Einsatz für meinen Erfolg waren eine Quelle der Motivation und Ermutigung während dieser Reise.

Ich bin auch dankbar für die Beratung und Betreuung durch Professor Silvia Miksch. Ihre Sachkenntnis und ihr Feedback waren für die Verfeinerung meiner Forschungsarbeit und deren Verwirklichung von wesentlicher Bedeutung, und ich bin dankbar für die Gelegenheit, mit ihr zusammenarbeiten zu können.

Außerdem möchte ich den Personen, die an der Evaluierung dieses Projekts teilgenommen haben, meinen aufrichtigen Dank aussprechen. Ihre Bereitschaft, sich die Zeit zu nehmen, an meinem Interview teilzunehmen und mir Feedback zu geben, war entscheidend für die Verfeinerung und Stärkung meiner Arbeit. Ihre detaillierten Kommentare und differenzierten Antworten auf meine Fragen waren für die Ausarbeitung meiner Argumente und die Verbesserung meiner Schlussfolgerungen von großem Wert.

Abschließend möchte ich meiner Familie und meinen Freunden für ihre Unterstützung und ständige Ermutigung danken, die während des gesamten Prozesses eine Quelle der Inspiration und Motivation für mich waren.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.



# Acknowledgements

I am deeply grateful to several individuals who provided support, guidance, and inspiration throughout my thesis process. First and foremost, I would like to thank Dr. Victor Schetinger and Dr. Johannes Schreiber for their invaluable contributions. Johannes' original idea for this project was the driving force behind my research, and Dr. Schetinger's continued support were instrumental in shaping the direction and scope of my study. Their feedback and commitment to my success have been a constant source of motivation and encouragement throughout this journey.

I am also grateful for the guidance and mentorship of Professor Silvia Miksch. Her expertise and feedback were essential in refining my research and bringing it to fruition, and I am grateful for the opportunity to have worked with her.

Furthermore, I would like to express my sincere gratitude to the individuals who participated in the evaluation of this project. Their willingness to take the time to participate in my interview and provide feedback was critical in refining and strengthening my work. Their detailed comments and nuanced responses to my questions were of immense value in shaping my arguments and improving my conclusions.

Finally, I would like to thank my family and friends for their support and constant encouragement, which were a source of inspiration and motivation throughout this process.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Kurzfassung

Der Ausbau der Windenergie auf globaler Ebene erfordert zuverlässige und aussagekräftige geografische Informationen über die Standorte von Windkraftanlagen. Studien haben gezeigt, dass man mithilfe von OpenStreetMap umfangreiche globale Daten über die Windinfrastruktur generieren kann, jedoch wurde bisher wenig Aufwand in die Darstellung und Analyse dieser Daten mittels visueller Werkzeuge gesteckt. Um eine präzise visuelle Untersuchung durchzuführen, ist es von großer Bedeutung, die passenden Parameter zur Charakterisierung von Windparks zu kennen und die geeigneten visuellen Kodierungen für die globale und lokale Analyse zu nutzen. Aus diesem Grund haben wir eine Designstudie durchgeführt, die zur Erstellung eines Datensatzes namens *Enriched Data of Wind Farms* (EDWin) sowie eines Prototyps für dessen interaktive Visualisierung führte. In einer Nutzer:innenstudie haben wir die Leistung des Prototyps bei der Überprüfung unbestätigter Behauptungen über Windparks aus der Literatur sowie bei der Identifizierung spezifischer Merkmale von Windparks durch vereinfachte visuelle Kodierungen evaluiert. Der Prototyp ermöglichte es den Nutzer:innen, die Aufgaben erfolgreich zu absolvieren, jedoch benötigten manche von ihnen Unterstützung durch die Befragende Person aufgrund des Bedarfs an einer verbesserten Funktionalität zur dynamischen Gruppierung. Darüber hinaus zeigten Interviews mit Expert:innen aus der Windenergiebranche, welche Merkmale für die Community von Bedeutung sind, um Windparks zu charakterisieren. Diese lassen sich in technische, zeitliche, Gelände- und Wettermerkmale unterteilen. Aufgrund der erfassten Merkmale konnten mehrere Erkenntnisse gewonnen werden, darunter die Feststellung, dass die weltweit vorherrschende Landbedeckung für die Installation von Windkraftanlagen landwirtschaftliche Flächen sind und dass die vorherrschende Landform flaches Gelände ist.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

The global expansion of wind energy requires robust and meaningful geographic information about its locations. Studies have shown how enriched global data on wind infrastructure can be generated using OpenStreetMap but have neglected to represent and make it analyzable using visual tools. For an accurate visual investigation, knowing which parameters can be used to characterize wind farms and which visual encodings are suitable for global and local analysis are essential. With this aim in mind, we conducted a design study that produced a dataset called the Enriched Data of Wind Farms (EDWin) and a prototype for its interactive visualization. Through a user study, we evaluated the tool's appropriateness for exploring unproven claims about wind farms from the literature and identifying specific wind farms characteristics through simplified visual encoding. The prototype enabled users to complete the tasks, but many needed help from the interviewer due to the need for an improved dynamic grouping functionality. Furthermore, interviews with wind energy experts revealed which features are relevant for the community to describe wind farms. They can be divided into technical, temporal, terrain, and weather characteristics. From those we have covered, several insights were generated, including that the worldwide predominant land cover for the installation of wind infrastructure is agricultural land and that the predominant landform is flat terrain.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Contents

<b>Kurzfassung</b>	<b>xi</b>
<b>Abstract</b>	<b>xiii</b>
<b>Contents</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation & Problem Statement . . . . .	1
1.2 Research Questions . . . . .	2
1.3 Structure of the Work & Main Contributions . . . . .	3
<b>2 Claims about wind farms</b>	<b>7</b>
<b>3 State of the Art</b>	<b>9</b>
3.1 Wind Energy in the Literature . . . . .	9
3.2 Data Models . . . . .	10
3.3 Wind Farm Visualizations . . . . .	12
<b>4 Methodology</b>	<b>17</b>
4.1 Design Study Framework . . . . .	17
4.2 Design Process . . . . .	19
<b>5 Data Collection</b>	<b>25</b>
5.1 Wind Turbines . . . . .	25
5.2 Processing . . . . .	29
5.3 Wind Farms . . . . .	32
5.4 Data Enrichment . . . . .	36
5.5 Final Data . . . . .	44
<b>6 Prototype Design and Implementation</b>	<b>45</b>
6.1 Design Decisions . . . . .	45
6.2 Prototype presentation . . . . .	47
6.3 Implementation . . . . .	53
	xv

<b>7</b>	<b>Evaluation</b>	<b>59</b>
7.1	Users . . . . .	59
7.2	Interview Structure . . . . .	60
7.3	Assignments & Results . . . . .	62
7.4	Summary & Discussion . . . . .	76
<b>8</b>	<b>Limitations &amp; Future Work</b>	<b>77</b>
<b>9</b>	<b>Conclusion</b>	<b>81</b>
9.1	Answers to the Research Questions . . . . .	81
9.2	Conclusion . . . . .	83
	<b>List of Figures</b>	<b>85</b>
	<b>List of Tables</b>	<b>87</b>
	<b>Bibliography</b>	<b>89</b>
	Appendix Prototype Manual . . . . .	95



# Introduction

## 1.1 Motivation & Problem Statement

Climate change is becoming increasingly noticeable and continues to progress, posing far-reaching risks to natural and human systems [1]. With the Paris Agreement, enacted in 2016, 196 parties of the international community have set a goal to counter the progression of global warming and limit it to below 2 degrees [2].

The energy sector is responsible for three-quarters of global greenhouse gas emissions, making it the largest polluter compared to other sectors [3]. Moreover, in 2022, there was a growth of 3% in energy demand, similar to the average growth of the last ten years, which suggests that energy demand in the future will still grow strongly [3]. At the same time, we are facing the first global energy crisis, which the International Energy Agency (IEA) calls a reminder of the fragile and unsustainable current energy systems [4]. Therefore, changes are urgently needed, and an acceleration of the growth of renewable energy infrastructure is essential to get on track towards fulfilling the goals and create a "more secure, sustainable and affordable energy system" [4].

On the global market, changes are becoming more noticeable in recent years. According to the "Renewables 2021: Analysis and forecast to 2026" report from the IEA, 2020 was the first year renewable energies generated more electricity than fossil fuels in the EU [5]. By 2026, renewable energies are expected to generate 37% of the world's electricity, becoming its largest contributor. The share of non-hydro energy sources is expected to double, reaching 18%, and will be led by wind and solar energy expansion [6] [7].

Among these, offshore wind farms will see the largest growth over the next five years by tripling their total capacity and thus generating 1.5% of energy from renewable sources in 2026. Onshore wind additions will also grow and be almost 25% higher on average in 2021-2026 than in 2015-2020 [6].

The increasing demand for renewable energy has led to the growth of wind energy as a significant power source. As a result, wind farms play a vital role in reducing global emissions and mitigating the effects of climate change. Understanding the distribution and characteristics of wind farms is important for researchers, policymakers, and the general public. However, information on a global scale is only available in summarized form, with installed wind infrastructure measured in terms of energy capacity, typically expressed in megawatts (MW) or gigawatts (GW). No information is included on where wind farms are distributed worldwide, nor their exact location, what they look like or what their characteristics are. Even in one of the most important sources about the situation and the progress of wind energy, the annual Global Wind Report of the Global Wind Energy Council (GWEC), such figures are hard to find [5].

The lack of information limits the land analysis on a global scale and the research potential for many application fields, such as estimates of capacity factors, power density, wake effects, and land requirements for wind energy; assessments of climate and health benefits of wind energy; the impact of wind turbines on local climate and surface temperature; analyses of public acceptance, noise pollution and land use preferences. Many more are elaborated in [8, p. 2]. Usually, studies that investigate these application fields are made on selected sample data, which leads to problems with its representativeness. For example, the study in [9] encountered this very problem when it sought to investigate the differences between onshore and offshore wind farms, where it concluded that wider spacing between the turbines and larger entities of installations characterizes offshore wind farms. However, the results were based on a dataset of 44 wind farms, which, the authors state, posed dangers to the validity of the statistical evidence [9]. The literature presents further claims about global wind infrastructure for which there is currently no possibility to prove them truthful. These claims are discussed in Chapter 2 and will be investigated within the thesis.

To analyze these aspects in the future, there is an urgent need for data and a tool that presents robust and meaningful geographical information on installed wind energy infrastructure globally. It may prove essential for future research and allow potential trade-offs for expanding the global energy grid in an effort to curb negative impacts, such as damaging biodiversity [7]. However, collecting and visualizing data on wind infrastructure can be challenging. With this thesis, we want to address the knowledge gaps in the domain and contribute with the help of visual analytics. We aimed to fill both gaps and have created a dataset we named the *Enriched Data of Wind Infrastructure* (EDWin) [10] and a tool for the exploratory visualization of it. The data contains the location of wind farms and turbines worldwide and has been augmented with characteristics obtained from the location.

### 1.2 Research Questions

The title of the thesis summarizes the problem we want to address, namely, what global wind farms look like. Obviously, there is no single answer to this question, as there are

thousands of wind farms, all of which are unique and look different from each other. The question "What do wind farms look like?" is a provocation on the current openness of the issue and our attempt to solve it through visual means. Thus, we conduct a design study to develop an interactive visualization that allows users to explore for themselves what wind farms look like, where they are and what their characteristics are. The resulting tool will enable experts to own the data and gather the insights they are interested in.

We have structured the problem into three research questions.

**RQ1** What features can be used to characterize wind farms worldwide, and what local patterns do they present?

**RQ2** What visualization technique is appropriate to explore the claims about wind farms found in the literature using the EDWin?

**RQ3** What visual encoding can accurately capture the main visual features of a wind farm (size, spacing, terrain, nearby infrastructure) and represent them for a [quick] clear visual comparison?

### 1.3 Structure of the Work & Main Contributions

Collecting and visualizing data on wind infrastructure is a challenging task. In this thesis, we present a study on designing and evaluating a prototype for the explorative visualization of global wind farm data, intending to improve our understanding of the potential for wind energy.

We begin by reviewing the current state of the art in wind farm data collection and visualization and describe the approach we take to design and evaluate a prototype for the explorative visualization of global wind farm data. Herefore, we distinguish between the visualization of individual wind turbines and the visualization of their grouping into wind farms.

One key aspect of our approach is using OpenStreetMap (OSM) as a source of location data for our prototype. We introduce our developed dataset, called EDWin (Enriched Data of Wind Farms) [10], based on OSM data and enriched with additional variables obtained from various databases. The decision for the enriching variables is based on the literature research and claims about wind farms, which we found in the literature but still need to be validated by representative data. The research questions will address the verification of these claims, with the aim of answering them using the developed tool.

We describe the process of cleaning and enriching the data and the unsupervised clustering and preliminary analysis we conducted to determine the appropriate parameterization to obtain wind farms. For this purpose, we develop a dedicated visualization tool that compares different clustering parameters on global hotspots of wind turbines. The tool is presented in Section 6.2.

The methodology of the design study combines the *Nine-stages framework* with two approaches for developing information visualization systems designs: the *Nested Model* and the *Design Triangle* [11] [12].

We then describe the resulting prototype, which allows users to filter wind farms and wind turbines according to various variables, visualize them on a map, view their frequency distribution over self-selected variables, and explore individual wind farms. We also present a specific visual encoding for wind farms that provides an intuitive understanding of their characteristics.

We evaluate the prototype through a user study with experts in wind energy, as well as with visualization and geography experts. Our evaluation finds that the prototype is useful for generating insights about the global wind infrastructure and fulfilling various assignments. However, we also identify limitations and potential areas for future work, including methods for estimating global turbine rotor sizes and power outputs, ideas for building a consistent database of wind infrastructure, and plans to make the prototype available to the public.

Finally, we analyze the results of our study and answer the research questions. We find that technical, temporal, terrain, and weather characteristics are all essential for describing wind farms. Insights we gained include that the predominant land cover for wind farm installation is agricultural land, and the predominant landform is flat terrain. We conclude by discussing the implications of these findings and our prototype's potential impact on wind energy research and policy.

The major achievements that the study has acquired are listed below:

- Analysis of wind farms and wind turbines visualizations in the literature and description of their advantages and disadvantages.
- A literature search for relevant characteristics for describing wind farms.
- A literature search for definitions of wind farms.
- The development of the EDWin dataset [10].
- A visualization tool for optimizing the parameterization to cluster wind turbines into wind farms using the DBSCAN algorithm based on random global hotspots of wind turbines.
- A detailed description of the methodology for conducting a design study that combines multiple approaches for designing interactive visualizations.
- A developed prototype that can be downloaded and used.
- The description of a user study to evaluate the prototype that combines qualitative and quantitative methods.

- A summary of important aspects gained through discussions with wind energy, visualization, and geography experts.
- A detailed description of the limitations and ideas for the continuation of this project.
- Through answering the research questions, a verification of the competence of the developed tool and the detection of important characteristics that describe wind farms in a general sense.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

## Claims about wind farms

In recent years, the global demand for clean and sustainable energy has led to an increase in the construction of wind farms. While wind energy has been hailed as a promising alternative to fossil fuels, it is important to ensure that the information used in research about wind infrastructures is accurate and reliable. Unfortunately, many claims related to wind farms lack substantial evidence and may be based on intuition rather than empirical data.

To address this issue, we conducted a literature review and identified three claims related to land use, size, and spacing of wind farms that require further investigation. These claims have been discussed in previous studies but have not been sufficiently validated with representative data. To address this gap, Research Question 2 aims to explore these claims using an exploratory visual analytics approach that can assist in assessing the validity of the claims based on available data. Our goal is to provide evidence-based insights that contribute to a more comprehensive understanding of wind infrastructures and help researchers and stakeholders make informed decisions about their development.

**Claim 1:** "Wind farms are often built on land that has already been impacted by land clearing and they coexist easily with other land uses (e.g., grazing, crops)." [13, p. 184-222]

**Claim 2:** "Wind farms comprise more than one turbine. In the United States and China there are some wind farms of several hundred turbines, but in Western Europe most wind farms comprise between 10 and 50 turbines." [14, p. 504]

**Claim 3:** "Offshore wind farms are characterized by turbines that are more widely spaced and they are much larger entities – generating far more power in aggregate than onshore counterparts." [9, p. 50]



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.



# State of the Art

This chapter will look at the current literature on wind farms. We want to find out for which areas of the domain the acquisition of global data would be significant. Finally, we will explore related work regarding dataset creation and wind farm visualizations. The most used search engines for literature research were the CatalogPlus from the TU Wien [15], ScienceDirect [16] and Google Scholar [17].

## 3.1 Wind Energy in the Literature

There has been a surge of literature on wind farms, with a focus on examining land suitability and construction opportunities for wind energy infrastructure, as in [18], [19], and [20]. These studies are applied locally and primarily concerned with the climate characteristics of the land area but neglect, due to limited capabilities, the assessment of already installed infrastructures in the place or similar regions. This limits the land analysis and disregards suitability on different aspects than the purely climatic one, like the socio-economic level, the impact on biodiversity in the region and on a global scale, and many other application fields [8, p. 2]:

- impact of wind turbines on local climate and surface temperature
- estimates of capacity factors, power density, wake effects, and land requirements for wind energy
- assessments of environmental and health benefits of wind energy
- analyses of public acceptance, noise pollution, land use preferences, and changes in property values associated with wind turbines
- estimates of the quality of wind resources in wind fleet

- mitigation of radar interference from wind turbines
- analyses of power grid impacts and requirements associated with high penetration of renewable energy
- geospatial analyses of the technical potential of renewable energy

To address these aspects, we need robust, up-to-date geographic information on the installed wind energy infrastructure worldwide. Acknowledging such information may prove essential for future research and enable potential trade-offs for maximizing the expansion of the global energy grid with an effort to limit negative impacts [7].

Another research confirming the need is the study "Do onshore and offshore wind farm development patterns differ" by Peter Enevoldsen and Scott Victor Valentine [9]. Among other questions, the researchers investigate whether there are differences in the shape and size of offshore and onshore wind farms and whether this results in differences in energy production. The study shows that offshore wind farms are characterized by the wider spacing between the turbines and larger entities of installations, resulting in much higher energy production than onshore parks. However, the authors performed the analysis on a relatively small dataset of 44 wind farms, which, they state, "posed dangers to the validity of the statistical evidence". Their results confirm the gap in the possibility of obtaining large-scale datasets.

A relevant domain for analyzing wind infrastructure is also spatial planning and designing policies for cultivating wind farms in a region. The Dissertation analysis "wind power deployment in urbanized regions" by Pia Nabielek [21] at the Technical University of Vienna provided exciting insights into the evaluation of the effectiveness of spatial planning concepts for the use of wind energy. She analyzed planning approaches for wind energy in three European urbanized regions: South Holland (Netherlands), Lower Austria (Austria), and East Flanders (Belgium). Each region set wind energy targets as part of its spatial planning and included zones for wind installations in its agenda a long time ago, which were considered appropriate at that time. However, since then, regions have failed to reflect on and evaluate the consequences of these decisions in light of technological and economic innovation and changing societal attitudes. Accordingly, policymakers may overlook new approaches that better fit the changing context to manage renewable energy deployment targets over time. The author mentions that as renewable energy continues to grow post-2020, innovations in planning practices along the way will be necessary to facilitate better exchanges between planning and implementation.

## 3.2 Data Models

In this section, we will examine relevant research related to our project. We will evaluate the current state of the art in obtaining global wind farm data, with a focus on using OpenStreetMap as a source.

### 3.2.1 Data

Recent studies have emphasized the importance of collecting granular data on the spatial distribution of wind infrastructure. One such study, "Spatial Distribution of Wind Turbines, Photovoltaic Field Systems, Bio-energy, and River Hydro Power Plants in Germany" [22], published in 2019, discusses how the expansion of renewable energy has led to conflicts with the goals of nature conservation and the acceptance of local communities. These conflicts include the reduction of agricultural land, plant degradation, visual impacts of land use, and other issues. The authors highlight the importance of researching the effects of renewable energy sources, such as wind and solar energy, on humans and the environment. Research can be conducted through case studies or close monitoring, but accurate location data is essential. To address this need, the authors compiled a publicly available dataset on renewable energy plants in Germany, including their spatial information and parameters such as capacity, total size, and commissioning year for wind turbines. This dataset was provided by the German authorities of the federal states and was collected and partially augmented by the authors. The dataset is useful for researchers and policymakers in Germany concerned with optimal spatial planning for allocating and expanding renewable energy infrastructures in compliance with societal and environmental requirements. However, such an undertaking is more difficult on a global scale as only a few countries provide accurate and publicly available national data.

The authors of the paper "Harmonized global datasets of wind and solar farm locations and power" [7], Dunnett et al., also recognized the issue of a lack of publicly available, global, and harmonized spatial data for wind and solar farms. Their publication was the main inspiration for our study's approach to generating and processing data.

They used the publicly available OpenStreetMap database to obtain spatial data on the location of infrastructures the wind energy. Through an analysis of key/value pairs, they identify the query that returns the most complete data, comparing a dataset containing 50 random wind turbines with known locations. They result with the combination of the tags `generator:source = wind` and `plant:source = wind` and get 305,306 wind point data and 1,899 wind polygons. In data processing, they classify according to the location, whether in the ocean or an urban area. Moreover, the authors perform the clustering algorithm DBSCAN to group the individual turbines into "wind farms" based on spatial proximity. The scale for spatial clustering is determined in advance using Ripley's Key function, which describes spatial characteristics of point patterns and indicates whether the turbines are clustered, dispersed, or single random points in space. With this project, the paper's authors are on a new path because for the term "wind farm", a definition in terms of spatial distances arises.

Furthermore, the authors estimated wind farm power capacity for part of the data solely from the number of turbines contained and the installation area. They used regression equations derived from three other independent datasets to get estimates based on the turbine density. According to the authors, these are the best possible estimates of wind capacity on a global scale at that time. Nevertheless, they mention that the reader should

take the estimates with caution: the datasets used for the regression equations come from very developed wind energy countries - USA, UK, and Germany - and thus, the wind farms are more likely to yield high power [7]. Hence, the importance of the location, number, and density of wind turbines in farms to estimate their capacities is evident. The farm layout and the average wind conditions at the installation site would also be factors worth considering for estimation models. An inappropriate layout design can be affected by wake loss, i.e. when turbines are located in the wind shadow of other turbines, and thus less wind is captured, and energy is lost [23]. To avoid related problems, there are many scientific articles dealing the optimization of wind farm layouts, coupled with the number of turbines and their density, making this a recognized problem in the wind energy community i.e. [24], [25].

### 3.3 Wind Farm Visualizations

We will investigate various techniques for visualizing wind data, including both individual wind farms and the distribution of multiple wind farms. We found it necessary to distinguish between those visualizing multiple wind farms and their distribution and those depicting individual wind farms and their layout, as we are interested in both.

#### Individual wind farms

The visualization of individual wind farms is common in publications dealing with wind farm layouts. They often show the wind farms in scatter plots, where the axes of the plots represent meters, e.g., Figure 3.1. Characteristics about the area of installation were found as shown in Figure 3.2, in which a heat plot indicates the elevation level of the terrain. The heat map can also be replaced by other characteristics, including climatic ones such as average wind speed.

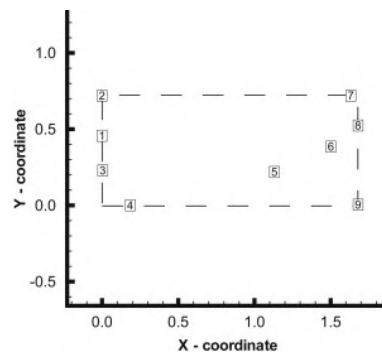


Figure 3.1: Wind farm layout in meters. Source: [26]

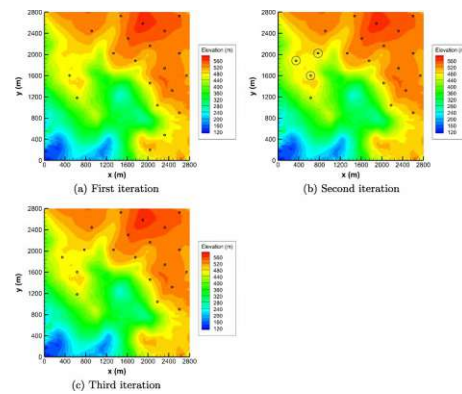


Figure 3.2: Wind farms layout visualization with heat map for elevation level. Source: [27]

Case studies on the analysis of already built wind farms use more site-specific visualizations, such as [28], where the authors determine important soil parameters for the foundation design of an offshore wind farm in Zhuanghe. They visualize the wind farm as Figure 3.3 shows.

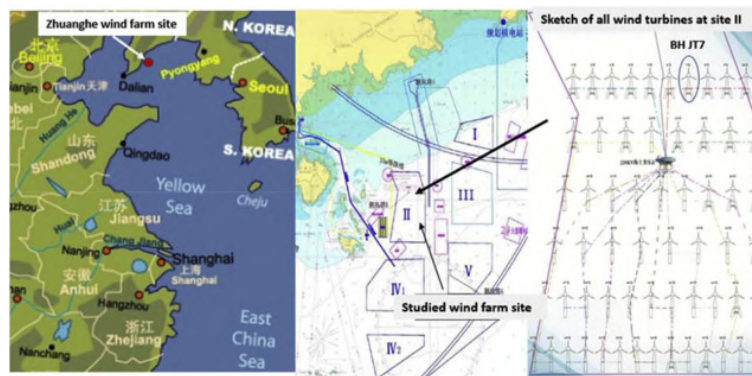


Figure 3.3: Wind farm visualization that combines location and layout. Source: [28]

This example is relevant to our study as the location of the wind farms plays a central role in our work.

### Overall wind farm distribution

The findings were limited regarding visualizations of overall wind farms and turbine distribution, for example, in countries, regions, or even globally. We found some publicly available websites that visualized country-based wind turbines, such as the USGS, which shows wind turbines in the USA. At further distances, the turbines are shown in a heat map; at closer zoom, the individual turbines are visible, where turbines belonging to the same farm have the same color. One can hover over the turbines to get information about



### 3. STATE OF THE ART

the year of construction, the capacity, the rotor diameter, the height and manufacturer and more. There is a screenshot in Figure 3.4 and, in the caption, the corresponding link.

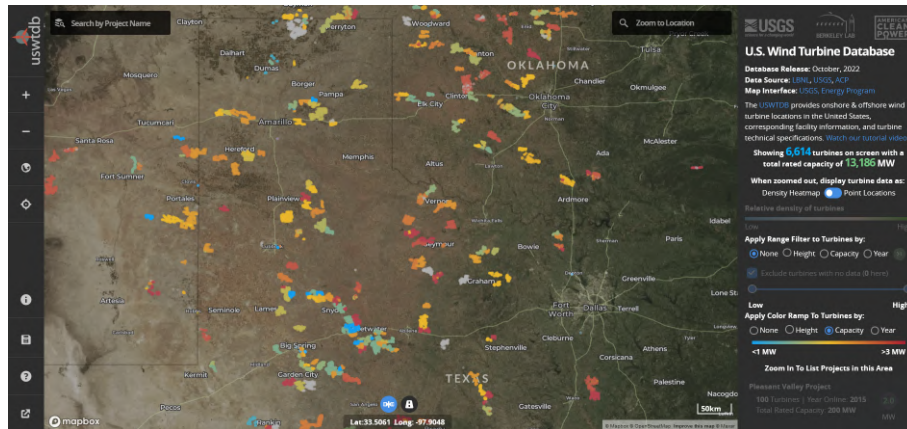


Figure 3.4: U.S. Wind Turbine Database Viewer. Source: [29]

This interactive visualization serves as a reference for our visualization since a map representation is also necessary to show the distribution of wind farms around the world. However, the given information about the turbines is very specific and would only be available for some parts of the world. Moreover, this visualization lacks information about other characteristics of the wind farms, for example on what kind of ground conditions or elevation they are located.

Another reference for geographical wind energy presentation, though again limited to data from the USA, is shown in Figure 3.5. It refers to all energy sources and aims to demonstrate their comparative deployment in the USA. However, wind infrastructure is not explicitly included in the visualization, instead aggregated capacities are shown in a bubble map.

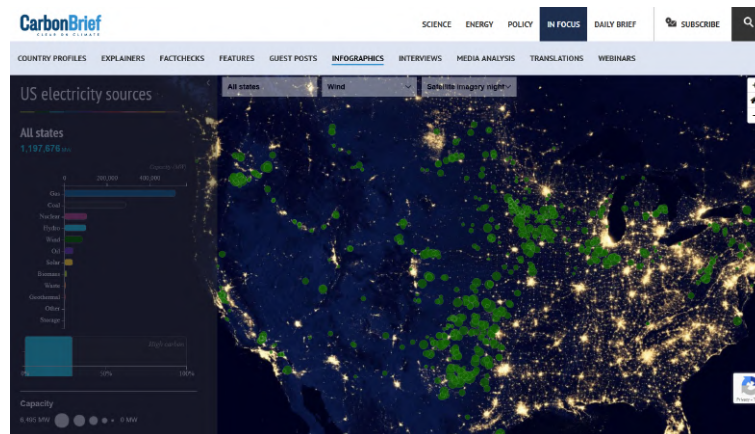


Figure 3.5: Mapped: How the US generates electricity. Source: [30]

To achieve the goals of the project, it was important to create visualizations with the exact location of each turbine and the characteristics associated with its location. While such visualizations were available for individual wind farms, we did not find corresponding visualizations for large-scale data, i.e., on a regional or global scale. In this respect, this project is a pioneer project.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.



# Methodology

This chapter presents the methods that we used to answer the research questions. It details how we followed the steps of the applied design study methodology and what additional methods we used to design and validate the visualizations. After defining a data-user-task model, specific requirements are established that will guide the data enrichment in Section 5.4 and the design decisions in Section 6.1.

## 4.1 Design Study Framework

We followed the nine-stages framework for conducting design studies by Sedlmair *et al.* [11]. To design and validate the visualizations we combined the *Nested Model* by Tamara Munzner [31] and the *Design Triangle* by Miksch and Aigner [12]. To gain access to expert knowledge and understand the domain problems, we collaborated intensely with an expert in the field, Dr Johannes Schreiber. He has a mechanical engineering background and specialized in wind energy, for which he held a chair at the Technical University of Munich (TUM). Now he has founded a startup and gives consultations on wind farm control, wake loss and energy production analysis.

We will first briefly introduce the frameworks for the design study and the visualizations. Then our use of them to develop graphic designs is described.

Sedlmair, Munzner, and Meyer developed the nine-stage framework as a basic methodology for conducting design studies. An overview of the framework is given in Figure 4.1. It consists of nine phases, divided into three categories: the **precondition phase** describes the requirements for conducting a design study; the **core phase** describes the steps of the actual design study, and the **analysis phase** the steps for interpreting the results retrospectively.

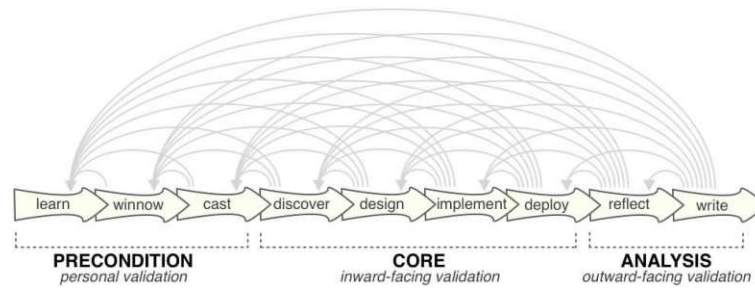


Figure 4.1: Nine-stages framework

The individual stages are carried out linearly but may overlap, and jumping back to already passed stages to edit them is often necessary. Validation occurs in every stage and is category-dependent. In the **precondition phase**, the validation is personal; in the **core phase**, it is inward-facing, i.e. evaluation of the outcomes together with the domain experts. In the **analysis phase**, it is outward-facing, i.e. justification of the results to the outside world. For each category, we will use Munzner's *Nested Model*, which gives appropriate methods for validation.

### Nested Model

Munzner introduced the *Nested Model* [31] for the development and validation of visualization designs. It consists of 4 layers:

- Characterization of tasks and data in the vocabulary of the domain problem.
- Mapping of those into abstract operations and datatypes.
- Designing visual encodings and interaction techniques.
- Creation of algorithms to efficiently execute visualization designs.

As Figure 4.3 shows, the layers are nested; Errors are propagated from the outer to the inner layers through the nesting of the system. To encounter this, validation methods are provided in each layer.

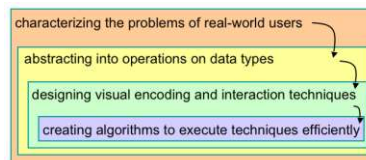


Figure 4.2: Nested Model

## Design Triangle

The *Design Triangle* is a framework for designing effective visual analytics solutions. It suggests that the design of a visual analytics system should consider three main aspects: the characteristics of the data, the user, and the tasks to be completed by the user. These three aspects form the three corners of the triangle.

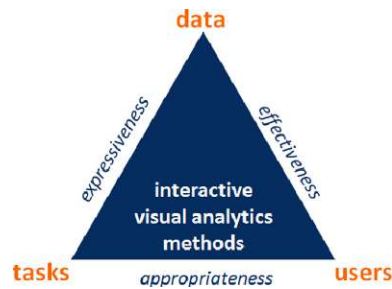


Figure 4.3: Design triangle

The characteristics of the data refer to the type, size, complexity, and structure of the data that will be visualized. Different data types and structures may require different visual encodings and analytical methods to convey the information they contain effectively. The user refers to the stakeholders using the visual analytics system. The design should consider the needs and capabilities of the user in order to support their tasks and decision-making processes effectively. The tasks refer to the specific goals and activities that the user will be performing with the visual analytics system. These may include exploration, analysis, comparison, or decision-making tasks.

Considering these three aspects together, the *Design Triangle* helps choose appropriate visual representations, analytical methods, and interaction techniques for the specific data, user, and tasks. This helps ensure that the visual analytics system is effective and efficient in helping the user to understand and gain insights from the data [12].

## 4.2 Design Process

In this section, we explain how we applied the methodology frameworks to our research.

### 4.2.1 Precondition phase

The first phase of the nine-stages framework consists of the *learn*, *winnow* and *cast* stages. We explored Literature on existing visualizations of wind energy data and described it in more detail in Section 3.3.

The Winnow phase for assigning the project's collaboration roles was predetermined as we had already collaborated with the domain expert for the interdisciplinary project. He provided interesting problems for visualizations for further projects and consequently originated this one as well.

The role of the domain expert was that of the front-line analyst, who analyses the data using the new visualization. The role of the gatekeeper was taken by the supervisor and co-supervisor, who were responsible for its authorization.

Three people were most involved in the project, the author of this thesis, taking the role of the visualization researcher, the domain expert and the co-supervisor. We decided to have regular meetings from the beginning, usually with all three people participating. Depending on the requirements, only the visualization researcher and the co-supervisor were involved, for example, when there were only bureaucratic. Everyone agreed to invest time over the next few months. After clarifying roles and time frames, the project was ready for the main phase of the project.

### 4.2.2 Core phase

The core phase consists of the *discover*, *design*, *implement* and *deploy* stages. The first three stages overlap with the *Nested Model* and will be explained in that context. All critical decisions for the design and creation of the visualizations were made here.

#### Problem characterization

The characterization of the domain problem followed from conversations with the domain expert and the statements found in the literature about wind farms, which are covered in Section 3.3. We have divided them into three types of domain problems we will refer to as P1, P2, and P2 corresponding to the research questions RQ1, RQ2, and RQ3.

**P1** is about the detection of global and local characteristics that can be used to describe wind farms. We address this issue with RQ1. It results from the discussions with domain experts and the insights they gain through using the developed solution.

**P2** is about making global wind farms' distribution, characteristics, and their relationships analyzable to explore the claims from Section 3.3. Thus, P2 corresponds to solving RQ2.

**P3** aims to develop farm-specific visual encodings to represent these characteristics and the appearance of P2 in a simplified and effective way. It should facilitate a comparison between a few wind farms by quickly looking at them in a visual sense to evaluate their size, spacing, terrain condition and nearby infrastructure. The solution to this corresponds to RQ3.

To fully characterize the scope of the application for which we are performing a design study, we sample a space of possible data, users, and tasks, thus coinciding with the *Design Triangle* framework, according to our problem characterization, the expectations of the domain expert, and our vision. This allows us understand requirements and instantiate prototypes over a design space.

**Data.** P2 and P3 build on the same data, yet in a different vocabulary. It is information about worldwide turbines and wind farms' location and characteristics. The relevant characteristics were derived from the statements about wind farms listed in the Research

Questions Section 1.2. For P2, the data is factual wind farm information about the *country, continent, land cover, landform, elevation, shape* and *turbine spacing*. For P3, it is the same information (with the addition of infrastructure near the wind farm) yet in the form of visual encoding in order for it to be legible from an image/map without explicit semantics.

**Users.** The following persons took the role of potential users of the visualization:

- A wind energy researcher who wants to enhance a publication with large-scale data
- An investor who wants to study wind farms in a particular area and evaluate construction possibilities
- A policy maker trying to create guidelines for better exchange of planning and implementation of wind farms
- A non-expert in the field who has seen a wind farm and wants to know its extent
- A student researching wind energy who needs information about the global wind resource infrastructure

**Tasks.** For P2 and P3, explicit tasks were defined, whose answers led the design of the visualizations. P2 tasks, take into consideration multiple wind farms:

1. Compare wind farms between countries and continents
2. Analyze how wind farms and turbines are distributed across land characteristics
3. Examine the distribution of spacing between turbines in wind farms
4. Find global hot-spots and cold-spots of wind farms and turbines
5. Discover different types of wind farm shapes

P3 tasks, take into consideration one or few wind farms. The tasks should be solved by mere observation.

Examine/compare the following visual information about the windfarm/s: (1) size, (2) distance between the turbines, (3) nearby infrastructure, (4) characteristics of the terrain

### Design requirements

The variables for enriching the dataset were selected depending on the selected claims in Chapter 2 and through consultation with the domain expert. We also weighed which variables would be feasible to obtain and which would probably be too complex for this research, resulting in the following variables:

## 4. METHODOLOGY

---

- Country
- Continent
- Land Cover
- Landform
- Shape
- Elevation
- Turbine Spacing
- Number of turbines (in wind farm)

To facilitate decision-making in the design process of the prototype, we formulated the following requirements based on the problem characterizations.

P2 specific:

- R1 The user should have the ability to view the distribution of global wind farms and turbines in the world, highlighting where there are global/regional hot spots and cold spots
- R2 It should be possible for the user to access detailed information about individual wind farms by clicking on them on the map.
- R3 The user should be able to view frequency distributions of wind farms and turbines over the selected variables
- R4 The user should be able to apply filters to the visualizations, such as displaying only a subset of installations or displaying only a subset of installations on the map, for all variables.

P3 specific:

- R5 Using the filter settings, the user should be able to generate a customizable grid display of a random sample of wind farms.
- R6 The user should have the ability to easily identify individual wind farms and their layout.
- R7 The user should be able to view the geographical characteristics of the land on which the wind farms are located.
- R8 The user should be able to see a representation of the spacing within the wind farms using a scale.

In the following Table 4.1, you can see a simplified representation of the methodologies we will use to answer the research questions, including the problem domains and the usage scenarios for Research Questions 2 and 3. The graph does not include all individual steps, but rather the most important ones with all the key points being represented.


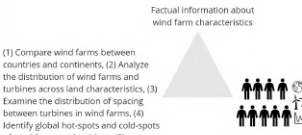

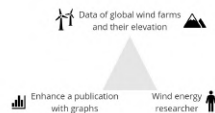
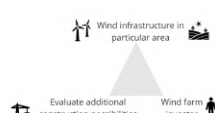
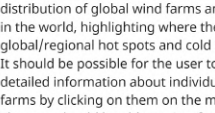
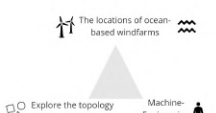
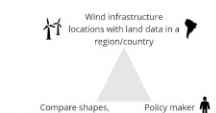
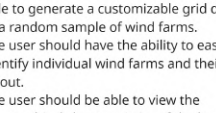
Research Questions	RQ 1 What features can be used to characterize wind farms worldwide, and what local patterns do they present?	RQ 2 What visualization techniques can be used to verify or reject claims about wind farms found in the literature using the EDWin?	RQ 3 What visual encoding can accurately capture the main visual features of a wind farm (size, spacing, terrain, nearby infrastructure) and represent them for a [quick] clear visual comparison?
Problem Domain	P1 Detect characteristics that can be used to describe wind farms.	P2 Make global wind farms' distribution, characteristics and their relationships analyzable to prove some claims.	P3 Develop farm-specific visual encodings to represent wind farm characteristics in a simplified and effective way.
Methodology	Expert Interviews 	Data-User-Task Factual information about wind farm characteristics 	Data-User-Task Visual information about wind farm characteristics 
Example Usage Scenarios		  	  
Requirements		<ol style="list-style-type: none"> <li>The user should have the ability to view the distribution of global wind farms and turbines in the world, highlighting where there are global/regional hot spots and cold spots</li> <li>It should be possible for the user to access detailed information about individual wind farms by clicking on them on the map.</li> <li>The user should be able to view frequency distributions of wind farms and turbines over the selected variables</li> <li>The user should be able to apply filters to the visualizations, such as displaying only a subset of installations or displaying only a subset of installations on the map, for all variables.</li> </ol>	<ol style="list-style-type: none"> <li>Using the filter settings, the user should be able to generate a customizable grid display of a random sample of wind farms.</li> <li>The user should have the ability to easily identify individual wind farms and their layout.</li> <li>The user should be able to view the geographical characteristics of the land on which the wind farms are located.</li> <li>The user should be able to see a representation of the spacing within the wind farms using a scale.</li> </ol>

Table 4.1: Simplified Methodology

### Evaluation Method

The concluding deployment stage concerns releasing a prototype and receiving feedback, in our context, the visualization evaluation, for which we will discuss the explicit structure

and results in the corresponding Chapter 4.2.2.

The main objective of the evaluation is to find out whether our solution supports users to solve tasks they were unable to solve before. For this purpose, we will conduct a *case study*, a common validation method of design studies. In this context, we will ask domain experts, visualization experts, and experts in other fields to solve tasks with real data for specific usage scenarios [11].

We followed the proposed methods of three publications to design an evaluation framework. The paper "Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau" by Leilani Battle and Jeffrey Heer was used as a basis to formulate and organize tasks for specific usage scenarios, which are closely linked to our data, users, and tasks. This paper is particularly useful because exploratory analysis is open-ended and can be challenging to be quantitatively evaluated [32]. To obtain a quantitative metric for the visualization quality, we followed the guidelines proposed in the publication "A Heuristic Approach to Value-Driven Evaluation of Visualizations" by Emily Wall et al. [33]. In addition, we also considered the criteria for rigor in visualization design study outlined in the paper "Criteria for rigor in visualization design study" by C. Plaisant, et al. These criteria are intended to help researchers evaluate the quality of visualization design studies and ensure that the design and results of these studies are transparent, valid, and generalizable [34].

### 4.2.3 Analysis Phase

The analysis phase includes the *reflecting* and *writing* phases and is done retrospectively. Thus, it is the process of writing this Master's thesis and reporting the conclusions. However, it already started at the beginning of the study by writing down notes, reflections, and procedures, which are now being elaborated.



# Data Collection

The following chapter first describes where the data originates from, how the individual turbine data was obtained and processed. It then describes how we defined wind farms in this work and how the wind farm data were acquired. We explain how the data was enriched with the variables decided with domain experts in Chapter 4. Finally, the chapter is concluded with a section on the quality of the data.

The resulting dataset called EDWin (Enriched data of Wind farms) is registered under the DOI <https://doi.org/10.5281/zenodo.7558885> [10]. The project under which the data collection and cleaning for EDWin was performed is available under the following GitHub repository: <https://github.com/marhal-beep/viz-windfarms-data-collection/>. This repository contains all the code used for collecting, cleaning, and processing the data to generate the final datasets.

## 5.1 Wind Turbines

The data for all wind turbines was extracted from OpenStreetMap (OSM) [35] with the help of the Overpass API [36]. The Overpass API is a read-only API that provides selected subsets of OSM map data. It works like a database over the Internet in that a client queries an API, and the API returns the dataset that answers the query request.

More information about OSM data and its structure will be given in the next subsection.

### Open Street Map

OpenStreetMap (OSM) is a free global, collaborative, open-source project for creating world maps and collecting worldwide geographical data. Users can upload vector data to the project's databases and edit it with the provided editors. Anyone that is registered can contribute by enriching or correcting the data [37].

The geographic data in OSM is distributed under a free license, called the Open Database License, and can be freely used for any purpose, including commercial purposes, as long as OSM and the contributors are cited as source [38].

A basic data element in OSM is either a node, a line or a relation. All elements can be assigned tags, which always consist of key-value pairs. Some attributes are not set directly by the users, but are entered and managed by the OSM system, including an ID that identifies the element. More information about the data structure can be found in the OSM Wiki [39].

### Wind turbines in OSM

Wind turbines in OSM [40] are linked to nodes and are extracted with the tag `generator:source = wind`. According to the official OSM wiki entry, the tag is obligatory, as is the tag `power=generator`, which is, however, implied by the first tag since OSM data is organized hierarchically. OSM wiki also specifies `generator:output:electricity=*` as obligatory, but we omitted it due to a radical reduction of resulting data points, likely caused by the inconsistency of OSM data tagging.

In [7], the authors also dropped the tag after performing a preliminary analysis to determine the most effective key-value pairs for extracting wind turbine data from OSM. They analyzed the tagging of 50 randomly selected wind installations with known locations and concluded that the tag `generator:source = wind` as a keyword would detect most of the target elements.

The data was extracted in Python with the package `request` [41] to send a query to the Overpass API. The explicit query is depicted below:

```
[out:json];
(node["generator:source"="wind"]);
out;
```

Applying the query to the entire globe resulted in the final raw dataset consisting of 362.325 data points.

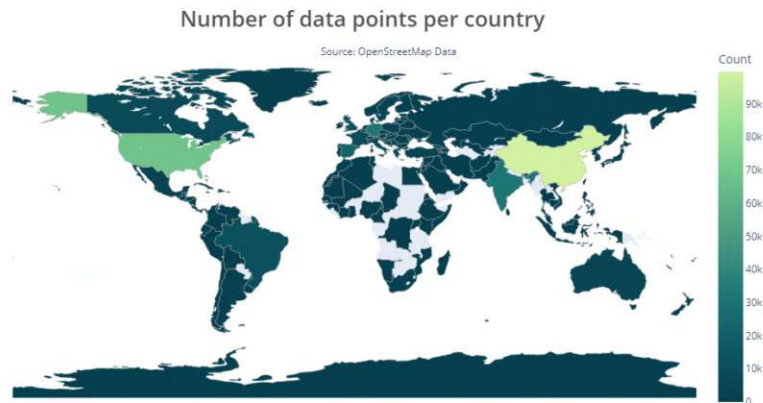


Figure 5.1: Number of OSM nodes per country

In addition to the locations in *longitude* and *latitude*, the key values *id* were extracted and used as identifiers throughout the project. According to the OSM wiki, we also downloaded other tags used in combination with wind turbines: power, hub height, manufacturer, and rotor size. Power refers to the capacity of wind energy production measured in different units, usually MW or GW. Height is the hub height of the rotor from ground level. Rotor size refers to the rotor's diameter, and manufacturer to the turbine supplier's name.

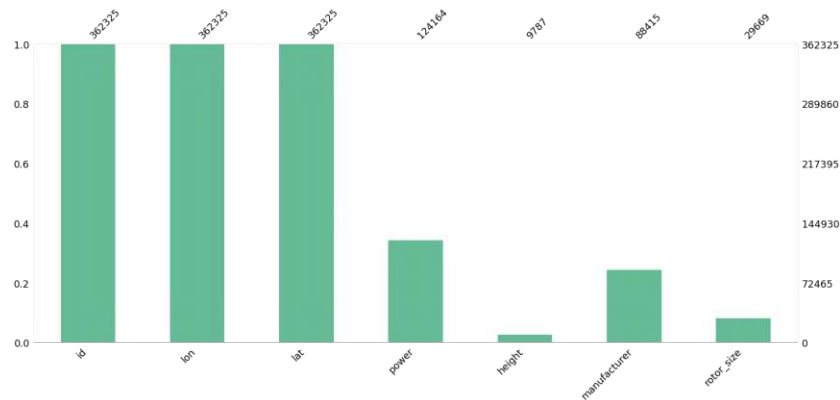


Figure 5.2: OSM data completeness

In Figure 5.2, the completeness of the data across the entire dataset is shown. The essential data about the location and the ID is complete. The additional tag variable with the fewest missing values is power capacity, and even that only achieves a completeness of 34.2%. The variables Hub height and rotor size have a totality of less than 20% and

## 5. DATA COLLECTION

manufacturer at about 25%. If one compares the data completeness further by country, one sees that this tag information is somewhat incomplete in most countries, especially China and India. However, as shown in Figure 5.1, these countries have the most data points. Given the high level of incompleteness, we decided not to include this information as a means of analysis and visualization.

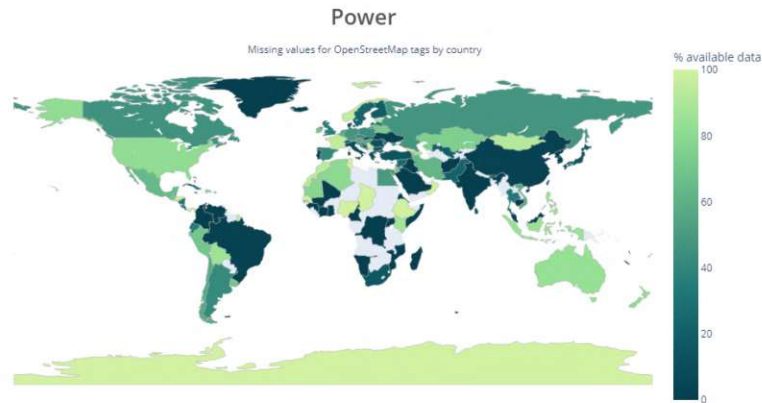


Figure 5.3: OSM power tag completeness by country

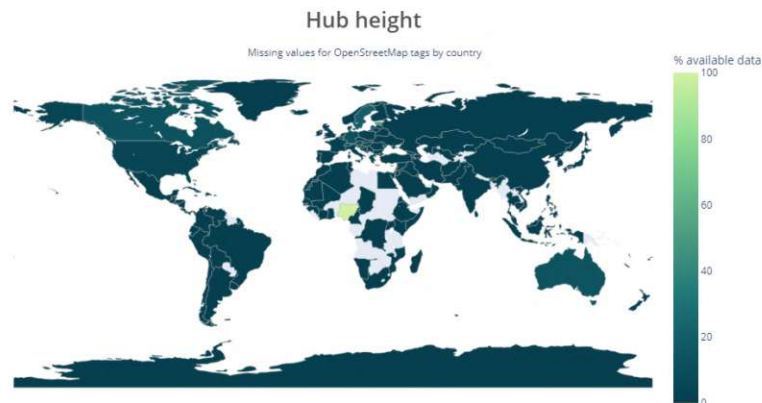


Figure 5.4: OSM hub height tag completeness by country

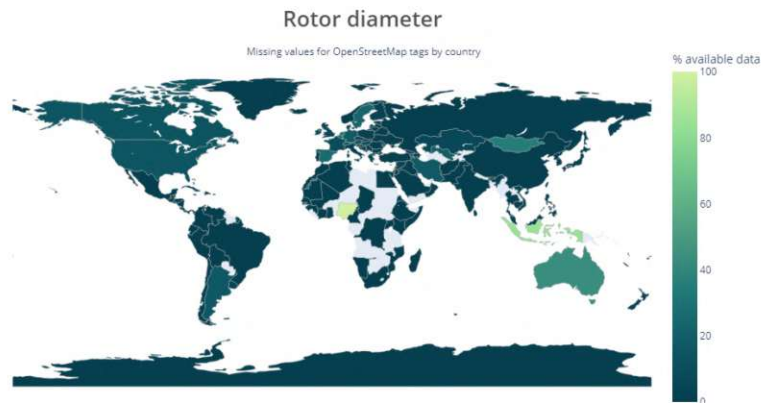


Figure 5.5: OSM rotor diameter tag completeness by country



Figure 5.6: OSM manufacturer tag completeness by country

## 5.2 Processing

Initial visualization and examination of the data using satellite imagery revealed that some data was inaccurate. Occasionally, there were markings where we could see no turbines on satellite images. Closer examination and research made us notice these were turbines

- not yet visible on the satellite images, but already built

## 5. DATA COLLECTION

---

- under construction
- planned but not yet built
- that had been dismantled

Since it would be difficult to check and correct the data in this regard automatically, we decided to leave the turbines in the dataset.

Further, we observed that some turbines were unrealistically close to others, sometimes less than a meter away. We assumed that the values were either entered twice and were incorrect or that one of the two points was a replacement turbine from an old turbine, and the outdated OSM point was still present. Examples can be seen in Figure 5.7.



Figure 5.7: Erroneous records for wind turbines

In order to analyze the situation in more detail, we decided to calculate the distance to the nearest neighbor for the turbines with closer neighbors. It was unnecessary to analyze all turbine spacings, as we were only interested in very close ones. Therefore, we used a density-based spatial clustering of applications with noise (DBSCAN) [42] algorithm to cluster the turbines into groups with a radius of 500 meters. This way, we would only have to consider turbines included in a group, and the distances would only have to be calculated between turbines within the same group. DBSCAN is described in more detail in the next section.

### 5.2.1 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that is used in our approach both for filtering erroneous data and, more importantly, to determine how to group wind turbines into wind farms (Section 5.3.1). The basic idea is that it clusters high-density points into the same groups and separates them from other groups through low-density areas. It requires two parameters:

$\epsilon$  : The range that specifies the neighborhoods. Two points are considered neighbors if the distance between them is less than or equal to  $\epsilon$ .

*minPts* : Minimum number of points to form a cluster.

The algorithm works as follows: Data points with *minPts* neighbors and distance less than  $\epsilon$  each represent core points used to form a cluster center. Data points that are not core points but are near a center are added to the cluster. Data points that do not belong to a cluster are considered noise points. More details on the algorithm can be found in [42].

### 5.2.2 Calculation of nearest Neighbour

We did the nearest neighbor calculation with the projected data to Eckert IV equal-area at 1 m resolution. Eckert IV was also used by Dunnett in [7] for clustering turbines in wind farms, so we adopted this projection.

To avoid turbines with no neighbors and to determine small groups of close turbines to make the computation less complex, we clustered them with a radius of  $\epsilon = 500$  meters using DBSCAN algorithm, available in Python's `Python` [43]. In each group, we then determined the nearest neighbor and the spacing for each turbine using the Nearest Neighbors [44] algorithm, available in `Python` as well. We used a `KDtree`, a balanced tree-based data structure, to store and organize the individual data points in order to minimize the distance calculations required [45].

### 5.2.3 Elimination of duplicates

As can be seen in the histogram of turbine spacings (Figure 5.8), there is a peak with more than 300 turbines at very small spacings of 0-1 meter. These were assumed to be duplicates. During a consultation with the domain expert, we discussed that the turbine spacings should generally be normally distributed, i.e., a few turbines with a small or big spacing, while most of them with a spacing between 200 and 400 meters.



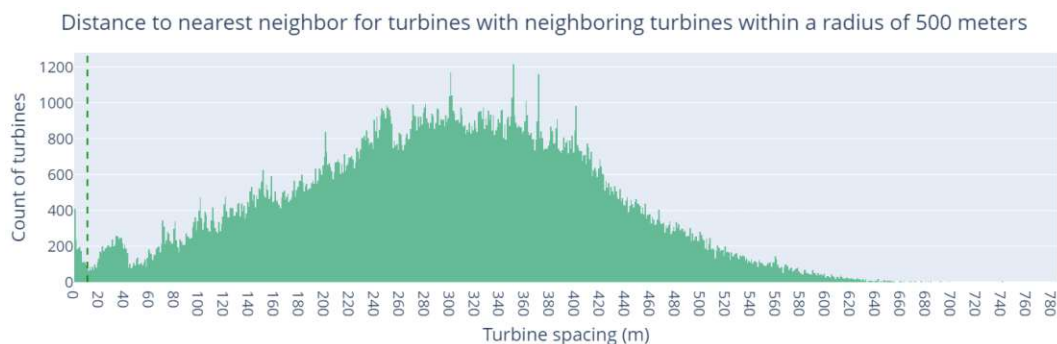


Figure 5.8: Turbine spacings for each turbine with neighbouring turbine

Consequently, we decided that up to the lowering of the first peak, we would include only one turbine for each neighboring turbine pair. The drop and rise after the first peak is marked with a green line at 10 meters. In other words, for all turbines less than 10 meters from the nearest turbine, one turbine was removed. The data was reduced by 2,378 turbines, resulting in a new dataset of 359,947.

### 5.3 Wind Farms

Obtaining wind farms from the turbine dataset required conducting a literature search on the definition of a wind farm. According to "Introduction to Wind Energy", a *wind farm* is defined as a group of wind turbines at the same location used to generate energy [46]. The "Dictionary of Energy" published in 2015, defines a *wind farm* as "An array or system of multiple wind turbines at a given site, used to capture wind energy for the production of bulk electricity for a grid. [So called because of the sense of "harvesting" wind as if it were a farm crop]" [47]. However, the question of what constitutes "the same location or site" and how many turbines form a "group" remains open.

The definitions show that the criteria used to define a wind farm are very open and often unclear. Factors such as property ownership, political, or geographical borders, turbine manufacturer, connection to the same energy grid, and operator/developer can all play a role in determining whether an accumulation of turbines constitutes a wind farm. For example, turbines may be considered part of a wind farm if they are all owned by the same entity, regardless of their location, or if they are located on the same property or land, regardless of who owns them. Similarly, a group of turbines located within the same municipality or county may be considered a wind farm, regardless of whether they are connected to the same energy grid. Additionally, the distance between turbines can also be a source of confusion, as some wind farms may have turbines located far apart but still connected to the same energy grid. All these factors can lead to confusion; thus, it is a subjective interpretation of different sources and usage contexts.

The question of "what do wind farms look like" should be preceded by the question of "what is a wind farm", as the lack of a clear answer to the latter question motivates the



approach taken in this work. Our thesis title, "what do wind farms look like", implies that we are attempting to understand and analyze the physical characteristics of wind farms. The lack of a clear definition motivates us to take an approach that aims to provide an evidential basis for specialists to formulate a proper definition.

Despite the confusion surrounding the definition of a wind farm, it was necessary to establish specific criteria to classify the turbines into wind farms for analysis. To accomplish this, we focused on a definition in terms of the parameters most relevant to DBSCAN clustering, namely spatial distances between turbines and the number of turbines within a farm.

### Definition

Our definition of *wind farm* is that a group of multiple wind turbines, with a minimum of two (i.e.  $minPts = 2$ ), located within a certain proximity of one another, constitutes a wind farm. The proximity is determined by the threshold distance,  $\epsilon$ , which we determined through analysis in the following section. We have disregarded factors such as the turbines being connected as a single electricity-producing power station and other factors, as we are primarily concerned with the spatial relationship between the turbines.

#### 5.3.1 Clustering

For clustering, all data points were first projected onto Eckert IV equal-area at 1 m resolution and then clustered using the algorithm DBSCAN already presented in Section 5.2.1. Two parameters are necessary for DBSCAN,  $minPts$ , the minimum number of points forming a cluster, and  $\epsilon$ , the radius for spatial clustering. The parameter for  $minPts$  was derived from the definition since "multiple wind turbines" implies more than one, i.e., at least two neighboring turbines are needed to form a wind farm; hence  $minPts = 2$ . Finding  $\epsilon$  will be discussed in more detail.

#### Finding the right radius for spatial clustering

The task of finding the  $\epsilon$  with which wind farms can spatially be identified was already addressed in [7]. As mentioned in Section 3.2, the authors performed an analysis to find the optimal neighbourhood radius for spatial clustering. They analysed the spatial characteristics of a known wind farm dataset, the United States Geological Survey (USGS) Wind Turbine Dataset (USWTD) [48], [49] to see if the wind farms are significantly clustered in space.

It resulted in an optimal value of  $\epsilon = 800$  meters. However, the dataset on which the analysis is based is from the USA, whereas the data on which the clustering was applied is global. We repeated the clustering from the paper using the same parameters, but the results were not very good. The problems we encountered are listed below. It suggests that the variety of types of wind farms were not taken into account using sole US wind farm data. In the example images, the red dots represent turbines, and the overlying blue polygon represents the clustered wind farm.

- The analysis does not include offshore wind farms, as there were hardly any offshore wind farms in the USA at the time [50]. Most offshore wind farms, however, are characterized by the fact that the turbines have a larger rotor diameter, and thus the distances between the turbines are larger. Figure 5.9 shows how the larger spacing led to the effect that either only certain sections of offshore wind farms were clustered (left image) or how far apart offshore turbines were not recognised as wind farm at all (right image).

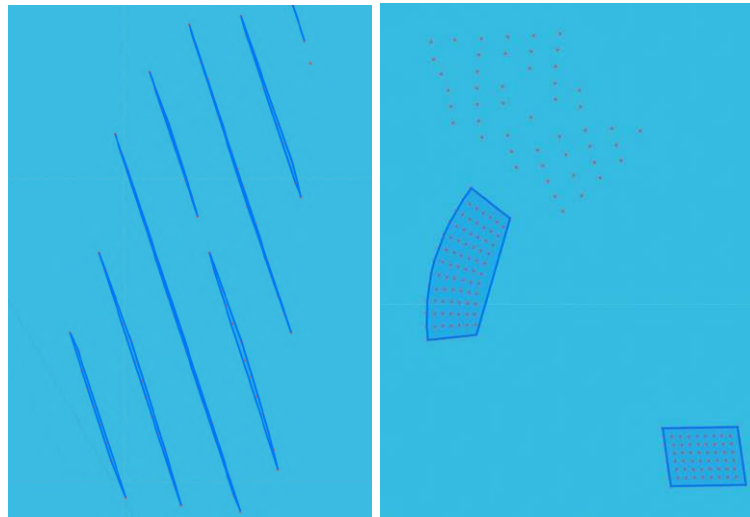


Figure 5.9: Offshore wind farms with only certain sections clustered (left) and wind farms not recognised as such at all (right) when using  $\epsilon = 800$  m

- There were also problems with clustering wind farms in other countries, e.g., China and India, which have farms with larger spacings than US farms. For example, Figure 5.10 and Figure 5.11 show farms where only parts are clustered, and others are not identified at all.



Figure 5.10: Erroneous clustering of wind farm in China with  $\epsilon = 800$  m

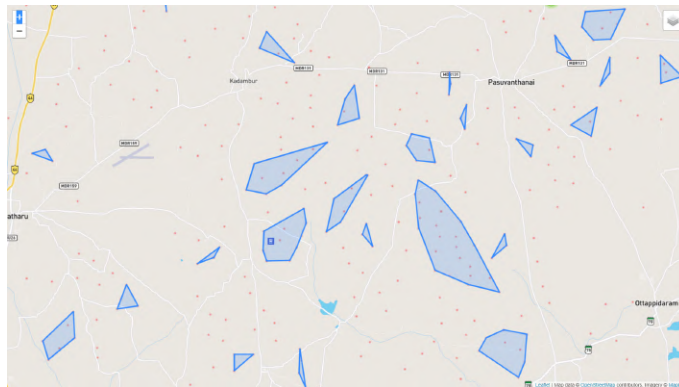


Figure 5.11: Erroneous clustering of wind farm in India with  $\epsilon = 800$  m

We decided not to adopt the  $\epsilon$  of 800 m but to determine a more accurate value.

Different parameters were tried, and the results were constantly visualised and assessed qualitatively. This gave some insights into the problem:

- Offshore and onshore wind farms have very different distances between turbines, making it challenging to capture them all with the same parameter.
- There are often significant differences in how countries build wind farms. In the US, for example, there are numerous wind farms with small distances between the turbines, i.e., small turbines, whereas such farms are hard to find in China.
- Similarly, there are differences in the construction of wind farms depending on the soil characteristics. Besides ocean-based wind farms that are very far apart, we found differences between farms located in urban areas or forests. This was also confirmed in the research [9].

The search for a suitable  $\epsilon$  proved to be a challenge that took much effort to solve in its entirety. After consultations with experts in the run-up and follow-up discussions, we found that the solution to this problem probably deserved its own research.

For this reason, we determined the clustering radius in a simplified way by mixing qualitative and quantitative assessments. We developed an interactive visualization tool, as shown in Figure 5.12, that allows users to explore different clustering levels of wind farms on random wind farm locations. Its primary purpose in this study was to aid in determining an adequate level for clustering, but it has the potential to be useful for a wide range of applications. It could help experts to visually explore and understand the effects of different clustering levels on wind farm data, and support future research by providing a previously unsupported capability.

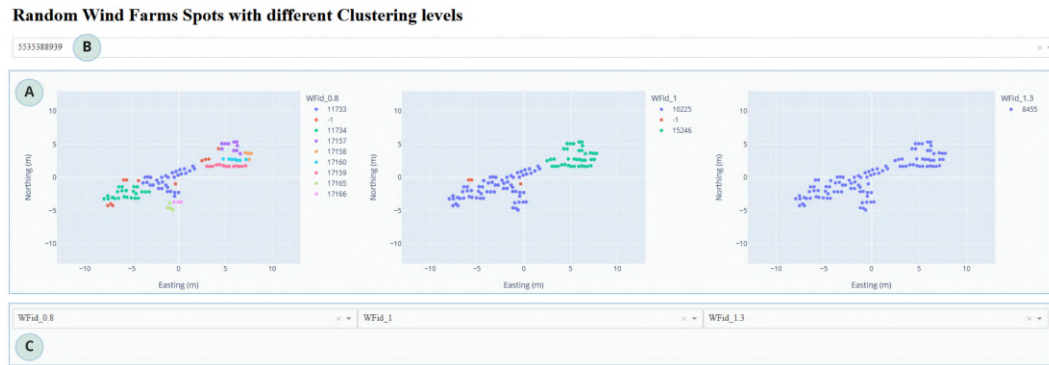


Figure 5.12: Tool to find spatial radius for clustering

The method used to determine  $\epsilon$  was by evaluating the clustering performance on a set of one hundred randomly selected global wind turbine sites, which are displayed individually in (A) and can be selected through (B). The locations of the turbines were projected using the Mollweide projection, which is an equal-area, pseudocylindrical map projection with a resolution of 1 m to their respective centers. This projection ensures that the actual areas and distances of the farms are accurately displayed in the scatter plots. The random sites were each clustered using  $\epsilon$  values ranging from 600 meters to 2 km at intervals of 100 meters. Three clustering levels are displayed at once, and the  $\epsilon$  level can be changed under (C) for each display individually. The naming of the clustering levels is interpreted as follows *WFid\_0.8* points to the clustering results for  $\epsilon = 800$  meters, *WFid\_1* means  $\epsilon = 1,000$  meters, and so on. In total, 15 clustering results are compared and evaluated with the help of domain experts. The best performing  $\epsilon$  was found to be 1.3 kilometres. Thus, all turbines that are less than 1.3 kilometres apart and have at least one neighbouring turbine within this distance are clustered as a wind farm. The result was 20.607 wind farms, with 353.782 turbines contained therein. 6.165 turbines were too far away from any other turbine and count as stand-alone turbines.

However, it should be noted that this value may not be considered "ideal" to define wind farms and the choice of  $\epsilon$  will depend on the specific goals of the analysis. It's important to remember that the purpose of this analysis is to find a working value of  $\epsilon$  for our specific research purpose, not to find the ideal value.

## 5.4 Data Enrichment

We enriched the data with the previously determined relevant variables defined in Section 4.2.2. These included the categorical variables:

- Country
- Continent

- Land Cover
- Landform
- Shape

and the numerical variables:

- Elevation
- Turbine Spacing i.e. Distance to the nearest next turbine
- Number of turbines i.e. wind farm size

The data for *Country*, *Continent*, *Land Cover*, *Landform*, *Elevation* and *Turbine spacing* were collected turbine specific and later added to the wind farm dataset in aggregated form. For the categorical variables, we took the modulus of the respective turbine values and, for the numerical variables, the average. For example, the *land cover* of a wind farm is the most common *land cover* value for the turbines located in it, and the *elevation* of a wind farm is the average *elevation* of its turbines.

The two variables, number of turbines (i.e. wind farm size) and wind farm shape (i.e. a rough shape of the wind farm), were obtained from the wind farms data and added to the turbine dataset. The collection of the individual variables will be discussed in more detail in the next sub-chapters.

A general structure of the data and their source is presented in Figure 5.13.

Wind turbine data: 359.947 entries, 12 columns

Column name	id	lon	lat	country	continent	land cover	landform	elevation	turbine spacing	WFid	number of turbines	shape
Source	Key value - OSM	OSM		reverse_geocoder, pyCountry Python libraries		Google Earth Engine geospatial datasets			calculation of the nearest neighbor for each turbine		wind farm specific data, same for all turbines in farm	

Wind farm data: 20.608 entries, 11 columns

Column name	WFid	lon	lat	country	continent	land cover	landform	elevation	turbine spacing	number of turbines	shape
Source	Key value - DBSCAN clustering	center of the wind farm		calculation of the modal value of the turbines in farm				calculation of the average value of the turbines in farm		number of turbines in farm	KNN unsupervised clustering of farm topology

turbine specific      wind farm specific

Figure 5.13: Data structure and source overview

## Country & Continent

The country was determined using the Python package *reverse\_geocoder* [51], which returns the official country name for latitude/longitude values. Then, with the help of *pyCountry* [52], an ISO database for the standards, the country name was translated into the continent.

### Land Cover, Landform & Elevation

This information was collected using Google Earth Engine [53], a visualizer of imagery data containing geo-information that features access to many geospatial datasets available through the Earth Engine Data Catalog [54].

The data was downloaded through the Earth Engine Code Editor [55], a web-based development environment for the *Earth Engine JavaScript API*. The code editor allows a user to analyze satellite imagery from the data catalog and download the values of datasets at specific locations. A different dataset was used for each of the required variables, giving a total of three datasets.

**Land Cover** was extracted from the 2020 "Copernicus Global Land Cover Layers-Collection 2" dataset [56], a global discrete land cover map at 100 m resolution. Table 5.1 shows the adapted naming of the discrete classes with their official meaning. Closed forests and open forests were combined for all leaf species.

Label	Description
Unknown	Unknown. No or not enough satellite data available.
Shrubs	Shrubs. Woody perennial plants with persistent and woody stems and without any defined main stem being less than 5 m tall. The shrub foliage can be either evergreen or deciduous.
Herbaceous vegetation	Herbaceous vegetation. Plants without persistent stem or shoots above ground and lacking definite firm structure. Tree and shrub cover is less than 10 %.
Agriculture	Cultivated and managed vegetation / agriculture. Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
Urban	Urban / built up. Land covered by buildings and other man-made structures.
Sparse vegetation	Bare / sparse vegetation. Lands with exposed soil, sand, or rocks and never has more than 10 % vegetated cover during any time of the year.
Snow and ice	Snow and ice. Lands under snow or ice cover throughout the year.
Permanent water bodies	Permanent water bodies. Lakes, reservoirs, and rivers. Can be either fresh or salt-water bodies.
Herbaceous wetland	Herbaceous wetland. Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present in either salt, brackish, or fresh water.
Moss and lichen	Moss and lichen.
Closed forest	Closed forest, Tree canopy >70 %
Open forest	Open forest. Top layer- trees 15-70 % and second layer- mixed of shrubs and grassland
Oceans and seas	Oceans, seas. Can be either fresh or salt-water bodies.

Table 5.1: Discrete Land Cover naming convention, Source: [57]



**Landform** was extracted from the 2015 dataset "Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning" [58]. The dataset includes a comprehensive classification of landforms based on hillslope position and dominant physical processes at 30m resolution, shown in Table 5.2. More detailed information on how the dataset was obtained is described in the paper. The levels for classifying the data correspond to the standard conventions for naming.

Hillslope position	ID	Class name	TPI	Slope (°)	CHILI
Summit	11	Peak/ridge warm	$(0.0 < mTPIs < 1.0)$ and $(30 < (E_o - E_n) < 300)$		Warm
Summit	12	Peak/ridge	$(0.0 < mTPIs < 1.0)$ and $(30 < (E_o - E_n) < 300)$		Neutral
Summit	13	Peak/ridge cool	$(0.0 < mTPIs < 1.0)$ and $(30 < (E_o - E_n) < 300)$		Cool
Summit	14	Mountain/divide*	$(0.0 < mTPIs < 1.0)$ and $((E_o - E_n) \geq 300)$		
Summit	15	Cliff		>50	
Upper slope	21	Upper slope warm	$(0.0 < mTPIs < 1.0)$ and $((E_o - E_n) \leq 30)$		Warm
Upper slope	22	Upper slope neutral	$(0.0 < mTPIs < 1.0)$ and $((E_o - E_n) \leq 30)$		Neutral
Upper slope	23	Upper slope cool	$(0.0 < mTPIs < 1.0)$ and $((E_o - E_n) \leq 30)$		Cool
Upper slope	24	Upper slope flat	$(0.0 < mTPIs < 1.0)$ and $((E_o - E_n) \leq 30)$	<2	
Lower slope	31	Lower slope warm	$(-0.75 < mTPIs < 0.0)$ and $((E_o - E_n) > -5)$		Warm
Lower slope	32	Lower slope neutral	$(-0.75 < mTPIs < 0.0)$ and $((E_o - E_n) > -5)$		Neutral
Lower slope	33	Lower slope cool	$(-0.75 < mTPIs < 0.0)$ and $((E_o - E_n) > -5)$		Cool
Lower slope	34	Lower slope flat	$(-0.75 < mTPIs < 0.0)$ and $((E_o - E_n) > -5)$	<2	
Valley bottom	41	Valley	$(mTPIs < -0.75)$		
Valley bottom	42	Valley (narrow)	$(mTPIs < -1.2)$ and $((E_o - E_n) \leq -5)$		

Classes were based on dominant hillslope position and defined using the topographic position index (TPI), slope, and continuous heat load index (CHILI).

\*Difference in elevation calculated at  $r = 2430$ , others are calculated at  $r = 810$ . ID is the unique identifier used to label each class in the landform dataset.

doi:10.1371/journal.pone.0143619.t001

Table 5.2: Discrete Landform naming convention, Source: [58]

For this dataset, data points on water were not classified, resulting in missing values in our turbines' data. To avoid the missing data, we created a new *landform* category, "waterbody". The points that were on "Oceans and seas" or "Permanent water bodies" in the *land cover* variable and were missing in the *landform* dataset were classified into the new variable "waterbody".

**Elevation** data were primarily extracted from the dataset "USGS EROS Archive - Digital Elevation - Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)" [59] from 2010, which includes global elevation data from several sources. Since the data was incomplete, the missing values were supplemented with elevation data from the Open-Elevation API, which can be found at the following website [60]. It is a free and open-source alternative to Google Elevation API. The documentation can be found at the following GitHub repository [61].

### Turbine Spacing

The turbine spacing was calculated as described in Section 5.2.2 for each turbine of the cleaned data. The data points were first clustered into groups with an  $\epsilon$  of 20 km, since with this radius no stand-alone turbine was obtained and missing values could be avoided.

For each of these groups, the turbine spacings were calculated using Nearest Neighbour [44].

### **Number of turbines**

The number of turbines variable is a wind farm specific variable and results from the count of turbines that are part of the wind farm based on the clustering results.

### **Shape**

The shape of the wind farms is another farm-specific variable, and it was the most challenging one to determine since it results from an unsupervised attempt to cluster the topology of the farms into groups of similar shapes through their image representation. Due to unsatisfactory results, we used only wind farms with more than five turbines for the clustering. Since there are no underlying group labels and it is not sure if ground truth labels exist, it took much effort to build and evaluate such a model. Despite the difficulties encountered, such as unsatisfactory results and the lack of similar attempts in the literature, we persisted in this effort as it is important to determine their overall appearance in order to understand "what windfarms look like". Further, the shape is a key aspect of our research, as it relates to the third research question and the encodings used to represent the wind farms. These decisions were taken to ensure that the representation of the wind farms is as clear and accurate as possible to achieve a good understanding of the overall appearance of wind farms.

The dataset to be clustered was derived from scatter plot images produced for each wind farm as shown in Figure 5.14. All wind farms were projected onto their respective centers using the Mollweide projection, a pseudo-cylindrical map projection that is true to the surface [62].



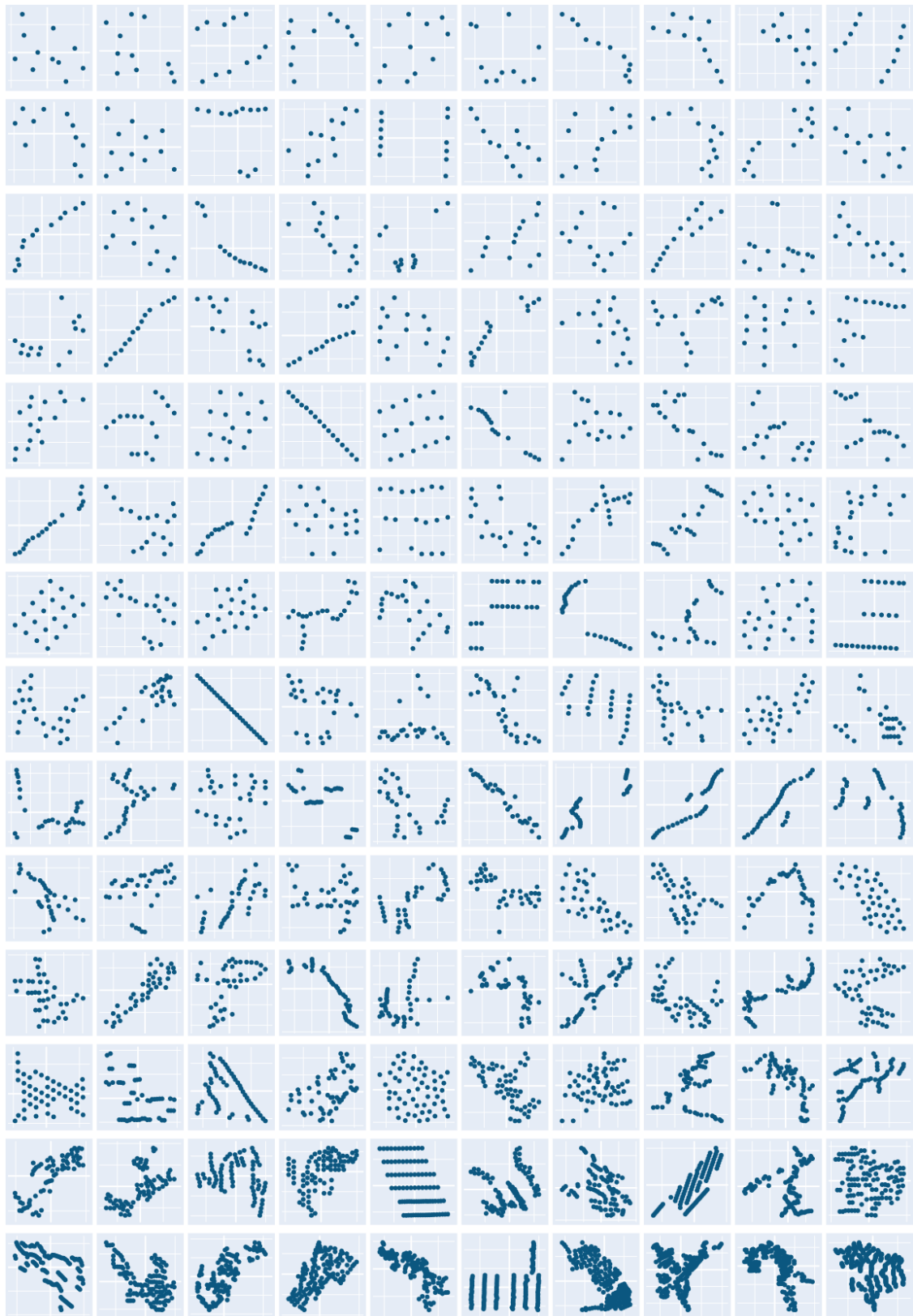


Figure 5.14: A glimpse into the "shapes" of wind farms

The methodology for unsupervised clustering was adapted from the study "Unsupervised Machine Learning Via Transfer Learning and k-Means Clustering to Classify Materials Image Data" by Ryan Cohn and Elizabeth Holm [63]. The authors explain how an unsupervised machine learning system for clustering images is built, used, and evaluated to classify unlabeled data. They use VGG16 [64] convolutional neural network pre-trained on the ImageNet dataset [65] to extract feature representation of the images. Afterwards, they apply principal component analysis to reduce the dimensions of the features while preserving relevant information and apply *k-means* to group the unlabeled data.

Methodologically, we performed the same steps as the authors:

1. Preprocessing: preparing the data for the VGG16 neural network.
2. Feature extraction: using VGG16 to create numerical representations of each image
3. Clustering: Labelling of data based on groups of similar information representations.

We implemented the clustering in Python. The images were created with Plotly [66], and the wind farm images were preprocessed with Keras [67] to prepare them for VGG16. We also used Keras to access the neural network. The obtained features were further processed with Python [43], and we determined its main components through PCA analysis. Fifty principal components accounted for 95% of the variance.

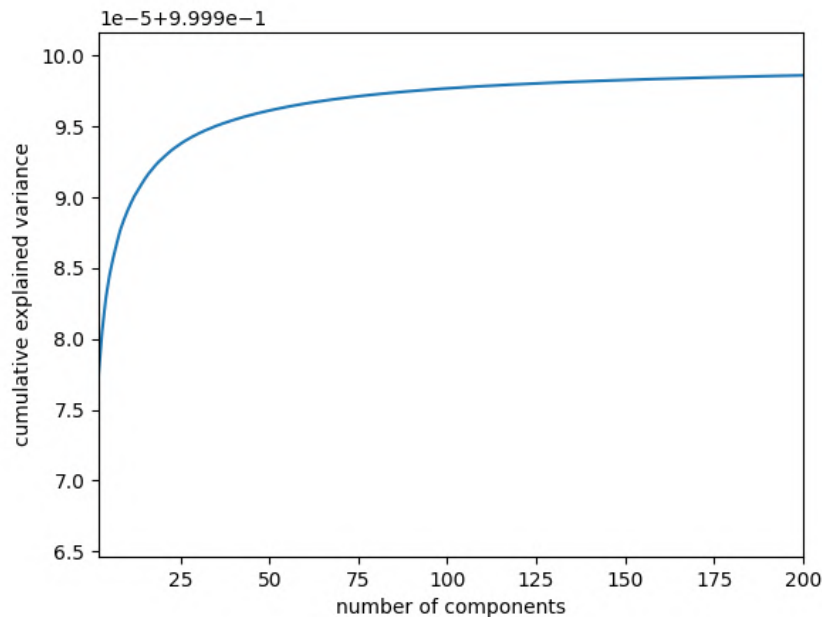


Figure 5.15: Explained variance by number of components

For finding similar groups, *k-means clustering* was used [68]. The algorithm works as follows:

1. Selection of  $k$ -points as initial centres of computation
2. Allocation of the data points to the different clusters based on the distance to the centres
3. Recalculation of the cluster centres
4. Repeat starting from step 2 - until the position of the centres does not change anymore

The selection of the components and the number of groups  $k$  was rather complex since there were no ground truth values, and we did not know the number of groups, so we could not optimize the model on the results for precision and recall.

Therefore, our selection was made using visualizations and qualitative assessments in combination with the Silhouette Score [69], a metric for the quality of clustering with unknown labels. The score's functionality is available in `PYTHON` as well. It is computed by the average intra-cluster distance  $a$  and the average nearest-cluster distance  $b$  for each data point with  $(b - a)/\max(a, b)$ . The overall Silhouette Score is the average score for all data points. The quality can be assessed with a score in the following ranges:

- Near 1: The data is well clustered
- Near 0: Many data points could belong to a different cluster
- Near -1: The data points are wrongly clustered

We calculated the Silhouette Score for different values of  $K$  and numbers of principal components but obtained worse results than we had expected. Most of the scores were very close to zero. The best score we got was 0.439 when we used 3 groups and 2 components. But when we looked at the groups visually, they seemed random. So we ended up focusing more on looking at the groups and deciding if they made sense, and we found that using more groups and components gave us better results.

We used  $K = 11$  and *Components* = 50, which achieved the visually strongest results. Several groups contained exclusively similar wind farms, such as line-shaped ones. Yet, other groups showed no similar types of wind farms. After consulting with the expert, we took those good groups and rated the others as unclustered. As a result, we obtained the following levels of wind farm shapes:

- Farms with less than 5 turbines (were not used for clustering)

- Lines
- Irregular lines
- Polygons
- Unclustered

## 5.5 Final Data

The resulting datasets are two, one for the global turbines and one for the wind farms we identified. The structure of the first and last three entries with missing values ratio per column can be seen in Figure 5.3 for the turbine dataset and Figure 5.4 for the wind farm dataset. The turbine dataset consists of 359,947 entries distributed among the 20,608 farms of the aggregated wind farm dataset.

row	id	lon	lat	Country	Continent	Land Cover	Landform	Elevation	Turbine Spacing	WFid	Number of turbines	Shape
1	12014318	-0.62960	52.36252	United Kingdom	Europe	Agriculture	Flat	930	1019	0	23	Polygon
2	21710548	0.71854	52.63015	United Kingdom	Europe	Agriculture	Flat	600	429	1	10	Not clustered
3	25105452	12.62040	55.69779	Denmark	Europe	Oceans and seas	Flat	10	125	2	7	Not clustered
359945	9999071537	109.02898	11.67856	Viet Nam	Asia	Agriculture	Flat	130	398	17298	47	Polygon
359946	9999071538	109.02707	11.68163	Viet Nam	Asia	Agriculture	Flat	240	454	17298	47	Polygon
359947	9999071539	109.02322	11.68319	Viet Nam	Asia	Agriculture	Flat	240	368	17298	47	Polygon
<b>Missing values</b>	0%	0%	0%	0.0072 %	0.0072 %	0.0089 %	0.1556 %	0 %	0 %	0 %	0 %	0 %

Table 5.3: Final wind turbines example data structure with ratio of missing data

row	WFid	lon	lat	Turbine Spacing	Elevation	Country	Continent	Land Cover	Landform	Number of turbines	Shape	popup
1	0	-0.65066	52.36485	407	83	United Kingdom	Europe	Agriculture	Flat	23	Polygon	Displayed
2	1	0.72542	52.62696	384	56	United Kingdom	Europe	Agriculture	Flat	10	Not clustered	Displayed
3	2	12.62637	55.69696	124	0	Denmark	Europe	Oceans and seas	Waterbody	7	Not clustered	Displayed
20605	20604	14.23328	52.55139	419	57	Germany	Europe	Agriculture	Flat	5	Less than 5 turbines	Displayed
20606	20605	108.89309	11.41953	571	33	Viet Nam	Asia	Agriculture	Flat	40	Polygon	Displayed
20607	20606	109.02134	11.66375	284	6	Viet Nam	Asia	Agriculture	Flat	10	Not clustered	Displayed
<b>Missing values</b>	0%	0%	0%	0%	0%	0.0008%	0.0008%	0.0011%	0.0114%	0%	0%	0%

Table 5.4: Final wind farms example data structure with ratio of missing data

The few missing values of three categorical values were filled with the value "Unknown", and the corresponding categorization was added to the variable.

### 5.5.1 Data Quality

Note that the data is not a one-to-one representation of the real values of the global wind infrastructure. In [7], the authors analyzed the issue of representing the real-world data with OSM data by regressing the raw numbers of wind features from OSM for each country to explain their respective reported wind capacities. They concluded: "the odds ratios for the national capacities suggest that the observed pattern is largely reflective of the true distribution of renewable infrastructure." [7, p. 9]. Hence, to the best of our knowledge, this is currently the most complete and accurate representation of the global wind infrastructure available.

# Prototype Design and Implementation

The prototype was designed and implemented according to the requirements and tasks specified in Section 4.2. The interactive visualization is available through the GitHub repository: <https://github.com/marhal-beep/viz-windfarms-dash>. This chapter presents the layout and explains the implementation of the single components.

## 6.1 Design Decisions

The visualization was designed in cooperation with the domain expert and the co-supervisor. It was based on the requirements elaborated from the data-user-task model in Section 4.2.2. Different visualization methods were discussed and validated using the *Nested Model* validation scheme [31] to keep the most successful ones ultimately. **R1** would quite obviously contain a map view in which all farms and turbines are displayed.

To satisfy **R1** we decided to create a map view displaying all wind farms and turbines. With over 20,000 points for the wind farms and 300,000 for the turbines, a simple scatter plot on a map would be unfeasible. From previous projects in which many points/polygons were to be mapped, we experienced that one of the most efficient ways to do this was by clustering the points.

To get the wind farm layouts and information about the turbines for **R2** we thought of two strategies: make the markers clickable and then display the wind farm as a popup or a similar way. Alternatively, to enable the user to change the clustering of the markers to represent either wind farms or turbines. This way, users could zoom in on a wind farm and then switch the view to wind turbines, at which point the individual turbines of the wind farm would be displayed. Additionally, details on demand about wind farms and turbines should be displayed via hover information or popup.

**R3** would be best represented by a simple histogram showing the number of wind farms/turbines on one axis and the selected variable on the other axis. With this simple representation and additional data filtering, all pending tasks could already be accomplished. The filtering of **R4** was, therefore, essential. It should control the results of both the map and frequency views. For example, users can display only wind farms on the map or the distribution of turbine spacing of wind farms on flat ground. The filter should also be responsible for the output of the visualization that satisfies requirements **R5-R8**, for which an explicit visual encoding was developed. A first rough draft can be seen in Figure 6.1. It consists of a scatter plot of the individual turbines, each represented by a clear point. The scales should be reflected by a grid placed over the wind farm. The grid should reproduce distances consistently, and the axes should be movable. The characteristics of land area, nearby infrastructure, rivers, mountains, and elevation should be represented by a specific map layer. Contour lines should depict elevations and mountains.

The inspiration for the matrix-like format to visualize multiple wind farms came from Figure 5.14, which made us learn early on that comparison between wind farms would be made easier this way.

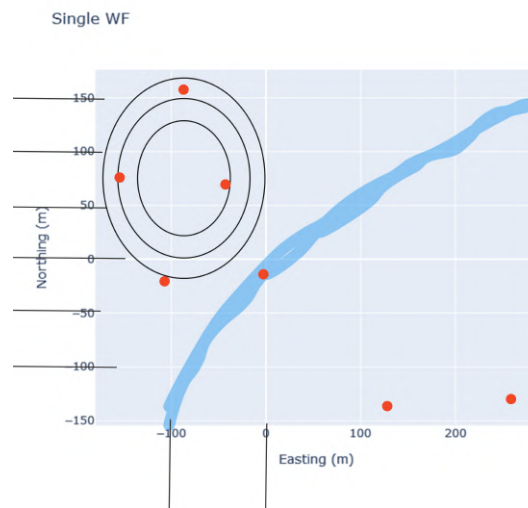


Figure 6.1: Draft of visual encoding for individual wind farm typology

This resulted in three different visualizations, all of which could be controlled by the same filter properties. They were ultimately designed into a common visualization consisting of three views.

## 6.2 Prototype presentation

The design and usage are presented with screenshots and descriptions of the user interface components. When the prototype is opened, it is displayed as in Figure 6.2.



Figure 6.2: Initial interface of the prototype

It has four primary functionalities: The filter and three data views. The filter can be opened via the button (A), and the views of the data are accessible through the tabs (B, C, D). When starting the application, the user interface looks like this: the filter is not visible, and the Map View is displayed automatically. At the top left corner (E), the visualization title is displayed, and in the upper margin (F), the number of farms and turbines resulting from the filter. This number always corresponds to the wind farms and turbines used to create the views. The functionalities will now be described in more detail.



## Filter

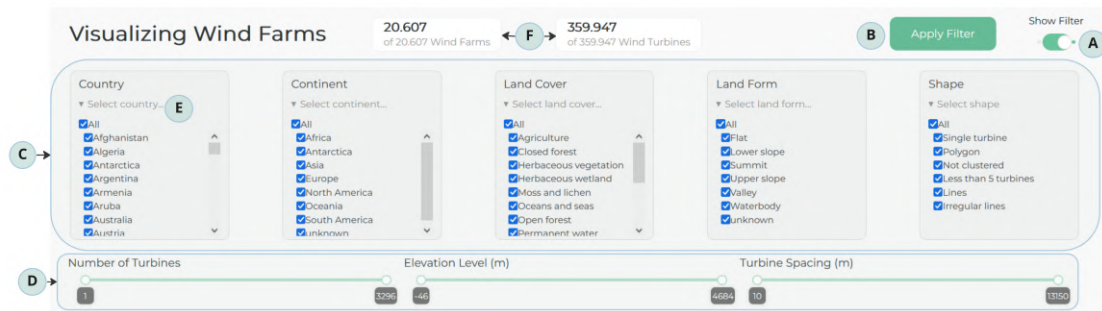


Figure 6.3: Upper section of interface with opened filter options

The filter (Figure 6.3) is a central component of the visualization, as it determines the results of the views. Only the data that fit the filter will be considered and displayed for the visualizations. It can be opened and closed via the button (A). It is applied to the data when the "Apply Filter" (B) button is pressed. Thus, the user can select all filters in advance, which are applied only once a final selection is made. If the data were filtered directly each time the filter was changed, overloading would occur quickly, and the application might crash. When the filter is opened, all wind farm characteristics are displayed as filterable variables. For categorical variables (C), a dropdown multi-selector menu (E) is available. These can be opened by clicking on "Select" with the variable's name, and then the corresponding variable's categories are displayed. There is a slider (D) for the numerical variables, where the upper and lower limits can be selected. The maximum and minimum values refer to the respective maxima and minima of the variables.

Filtering works differently for farms and turbines. For farms, the aggregated values are filtered; for turbines, the individual turbine characteristics are filtered. (F) shows how many turbines and farms apply to the filters. In Figure 6.4, only wind farms located on the *land cover* "Oceans and Seas" are displayed. The total number of wind farms and turbines shows that only 554 farms and 10.335 turbines are now being used for the visualizations. Likewise, only these are displayed in the initial map view.





Figure 6.4: Interface with activated filter only showing wind farms on oceans and closed filter options

### Map view

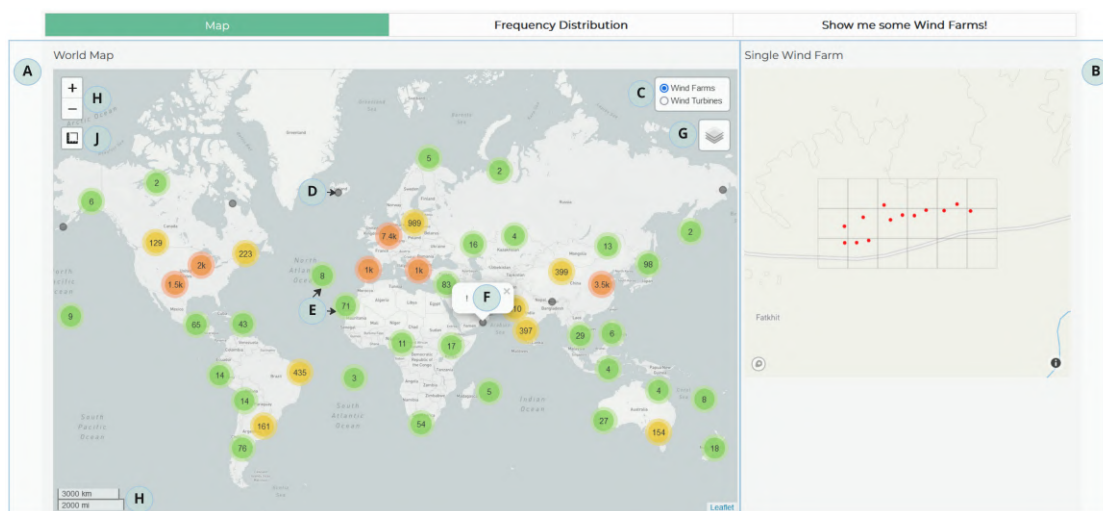


Figure 6.5: Map view interface: initial view

The map in Figure 6.5 is the default view when opening the tool. It gives an overview of where wind farms are distributed in the world (A) and allows to view the topology of individual wind farms (B). Under (C), one can select whether farms or individual turbines should be displayed on the map. Then the clustered markers (D) represent one of the datasets. The markers are clustered with a radius of 100 pixels and are recognizable by the colored cluster nodes (E). The cluster nodes indicate how many markers belong to

the cluster. The color of the cluster markers changes accordingly and has the following color encoding:

marker count	color
< 100 markers	green
100 - 1000 markers	yellow
> 1000 markers	red

Table 6.1: Color encoding for marker cluster on map view

Clicking on the cluster node will zoom into it. At higher zoom levels, more and more individual markers are displayed. If the markers are selected to be wind farms, the plot on the right side is displayed by clicking on the individual markers. The currently clicked marker is indicated by a popup (F). At (G), different layers for the map can be selected; the choices are the standard maps from Mapbox in *Light*, *Dark*, *Satellite*, and *Outdoors*. Zoom buttons are located at (H); one can either click on them or use ctrl+scrolling on the map for zooming. Below is a measuring instrument (J), which allows the user to measure distances and areas on the map. The map scale is displayed in the lower left corner (K). The plot of individual wind farms in (B) will be discussed in more detail in View 3, as they use the same encoding.

Frequency distribution view

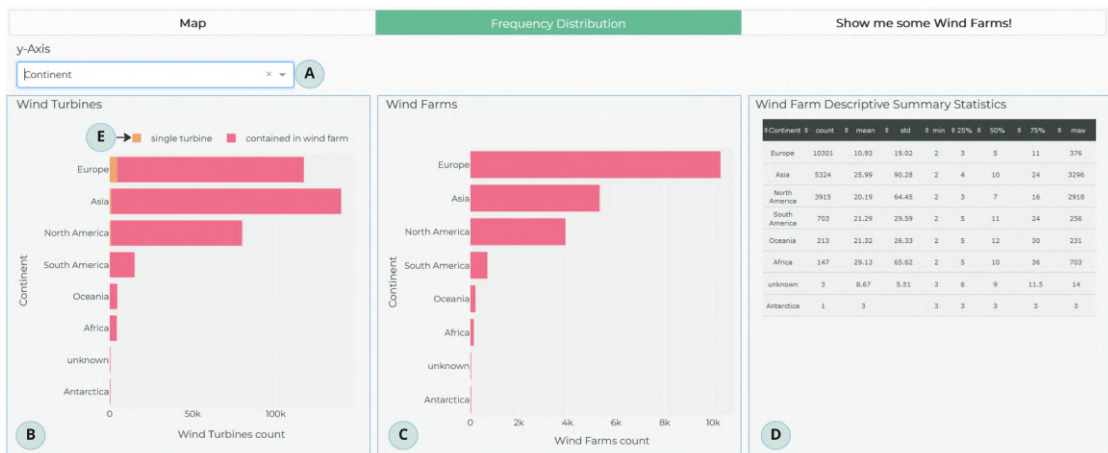


Figure 6.6: Frequency distribution view interface: Bar charts for categorical y-axis

The second is the frequency distribution view, intended to show how the frequencies of farms and turbines vary over a specific variable that can be selected via (A). The available options here correspond to the selected characteristics that describe the data and are also used for filtering. The view works differently depending on whether the selected variable (A) is categorical or numerical.

The categorical case is shown in Figure 6.6. It results in a bar chart for the individual turbines (B) and the wind farms (C). Table (D) gives a more detailed understanding of the wind farms' size in the respective variable categories. For each category, there is a summary statistics distribution of the wind farm size (number of turbines). The summary statistics include the number of wind farms in the category, the average, the standard deviation, minimum, maximum, and quantile distribution of the wind farm sizes.



Figure 6.7: Frequency distribution view interface: Histogram for numerical y-axis

The case where the variable (A) is numerical is shown in Figure 6.7. Here, a histogram is displayed instead of a bar chart for (B) and (C). Then, table (D) contains the summary statistics about the variable (A) itself. Here the summary statistics represent the number of wind farms, the mean value of the variable (A), and its standard deviation minimum, maximum, and quantile distribution.

### Random wind farms view

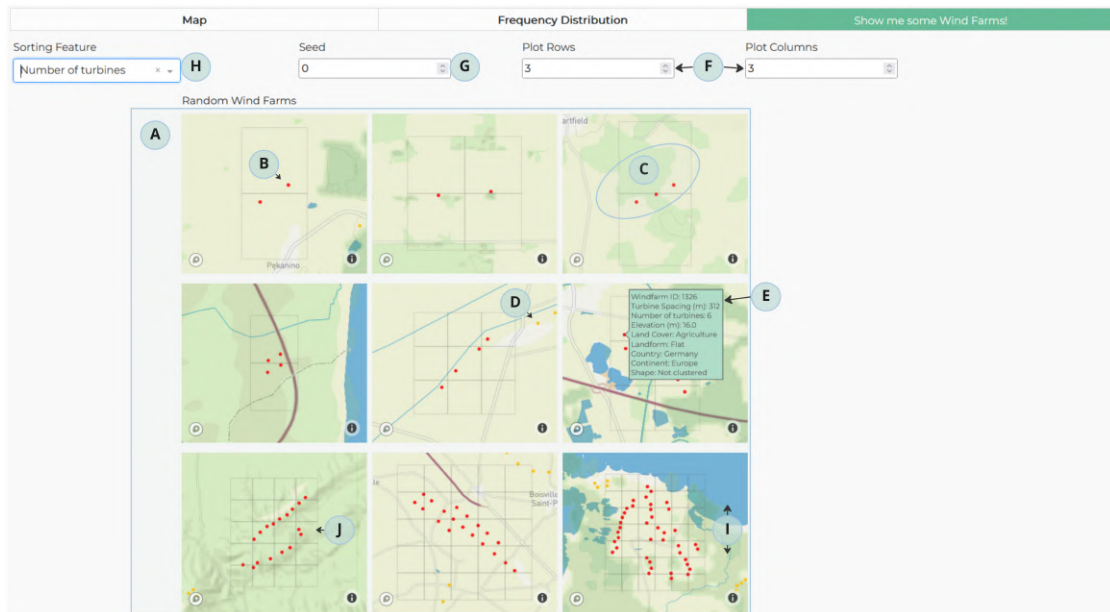


Figure 6.8: Random wind farms view interface

The third view shown in Figure 6.8 is meant to show random wind farms and give an idea about their installation area and layout. The wind farms are randomly selected from a pool of farms that match certain filters, and then plotted in a grid format similar to a poster (A). Each individual plot within the poster represents a single wind farm, where the single red dots (B) represent turbines included in the windfarm, the total cluster of red dots (C) represent the whole wind farm, and the orange dots (D) represent nearby turbines that are not part of the wind farm. When a user hovers over a dot, they will see tooltip information (E) specific to that turbine.

The number of wind farms shown can be adjusted by changing the number of rows and columns (F) of the poster. The user also has the option to change the seed (G), which will generate a new random sample of wind farms to display. The seed has the additional purpose of making the posters reproducible, meaning that with the same seed and filter settings, it will always be the same wind farm sample.

The wind farms can be sorted by a characteristic (H), for which the options again correspond to the filter features. In the Figure, for example, the wind farms are filtered by "number of turbines"; thus, the wind farms are shown in ascending order of size. The individual wind farm plots have a specific visual encoding composed of a map layer (I) and a grid (J). The custom map layer is designed to make it easier for users to see what *land cover* and *landforms* the farms are on, what infrastructure is nearby, and at what *elevation* they are located.

We used the following color encodings to illustrate these elements on the map:










Map element	Color
Oceans and seas, Permanent water bodies	
Herbaceous vegetation, Herbaceous wetland, Moss and lichen, Shrubs	
Closed forest, Open forest	
Agriculture	
Sparse vegetation	
Snow and ice	
Urban, Buildings	
Highways	
Elevation contour line	

Table 6.2: Color encoding of map elements used for individual wind farm encoding on "Random wind farms view"

In the overlaying grids, each line of a small square has a length of one kilometre. Its purpose is to make it easier and faster for the user to see the extent of the wind farm and the distances between the turbines. The visual encoding used were chosen based on the design requirements set set for P3 in Section 4.2.2. These design requirements likely outline the specific information and functionality that the visual representation needs, and the visual encoding is chosen to best convey that information in a clear and effective manner.

The visual in Figure 6.8 provides a clear representations of wind farms, including nearby infrastructure and the terrain on which the turbines are installed. The map layer allows to see roads or highways near the farms, as well as the type of terrain, whether they are on a mountain, in the ocean or on agricultural land. By comparing different wind farms, the user can easily identify the key differences between them, such as the size of the farm, its proximity to a highway, or the type of terrain it is located on. The user can also see the density of turbines on the farm. Overall, the goal of the visual encoding is to provide a clear and intuitive representation of the wind farm, allowing to quickly and easily understand key characteristics and differences at a glance.

## 6.3 Implementation

This section explains the technical details for implementing the developed designs.

### 6.3.1 Technologies

The prototype was written mainly in Python [70] and JavaScript [71], and interface aesthetics were done in CSS3 [72]. It was deployed online with Heroku [73] to give users easy and quick access during the evaluation phase. The website was accessible at

<https://viz-windfarms.herokuapp.com/>, but unfortunately, it has been taken down due to the removal of Heroku's free product plans.

The prototype was primarily built using the following libraries, which will be discussed in more detail in the following paragraphs.

- Dash
- Plotly
- Leaflet
- Mapbox
- Turf

### Dash

The prototype framework was implemented using the open-source library `Plotly Dash` [74], one of Python's leading user interface libraries. It allows the development of reactive and web-based analytical applications. The web applications run on `Flask` and communicate with `JSON` packages through `HTTP` requests. The front end of `Dash` renders the components in `react.js` [75].

`Dash` delivers a framework for integrating `plotly` visualizations into the dashboard. The applications are made reactive through the provision of so-called "callbacks functions", decorators consisting of the callback (the input and output variables) and a function that takes the input of the callback and returns any desired output, e.g., a visualization, data, and text. The outputs are bound to the `Dash` components, which are translated `HTML` components that build the application. The input elements are also passed through `Dash` components, e.g., a dropdown selector or a slider. Callbacks are executed every time an input variable changes [75].

### Plotly

`Plotly` [66] is a graphing library for Python that creates browser-based visualizations and is ideal for integrating graphs into `HTML` files or `Dash` applications. The visualizations are interactive, and over 30 plot methods produce different graphs [75].

### Mapbox

`Mapbox` [76] is a provider of interactive online maps for websites and applications. `Mapbox` provides the *Mapbox Studio web application*, a tool to design map layers, enrich them with data and retrieve them in different operating systems. `Plotly` has built-in functions for `Mapbox` and allows the easy utilization of previously designed `Mapbox` layers. The built-in functions include scatter, line, choropleth, or density plots. For this project, we mainly used the function `scattermapbox`, which plots points on a `Mapbox`



layer. It was used for the visualizations that required a specific visual encoding of the wind farms, relevant for RQ3 [77].

### Dash Leaflet

Dash Leaflet is based on the open-source library `leaflet.js` and is a wrapper for Dash applications. It provides interactive online maps and various mapping functionalities. We mainly used it to create the initial map view. The large number of data points made it appropriate to use the leaflet mapping function `supercluster`, which allows efficient spatial point clustering within a given radius [78].

### Turf

`Turf.js` is an open-source JavaScript library that enables spatial analysis with modular operations. It can be executed in any browser and used with mapping libraries such as Mapbox or Leaflet [79].

#### 6.3.2 Architecture

For the implementation of the prototype, a typical client-server architecture was used. Most of the functionalities are loaded on the server and rendered in the browser. However, due to problems with the time required to filter the data, we decided to move the data filtering to the client using Dash's `clientside-callbacks`. They run directly in the browser and do not need to send a request to Dash. They are helpful when a large amount of data needs to be sent back and forth, when callbacks are called often, or when the callback is part of a chain that requires multiple round-trips [80]. The filtering met these requirements since the data was relatively large at about 43 MB. The function of the `clientside` callback is written in JavaScript in the `assets` folder.

The input and output modality remains the same, the input values must be passed to the JavaScript function, and the outputs are returned as components. The filter callback is called when the "apply filter" button is clicked. The complete data and the current filter's state are then the callback's input. In order to use data as input, Dash requires that it is also stored in the browser. This is done using the `dcc.Store` function of Dash core components. The whole data is loaded into the browser at the beginning and is filtered by the `clientside-callback`. The output is the indexes of the filtered data, which are loaded into a `dcc.Store` and, from there, called by the server-side callbacks to create the visualizations. The transfer between Dash and the browser is the size of the indexes, i.e. the key variables of the respective wind farm and wind turbines dataset, which is about 4MB for the complete data.

Another functionality that moved client-side was the opening and closing of the filter. The other components and callbacks were left on the server side. A simplified architecture can be seen in Figure 6.9.

## 6. PROTOTYPE DESIGN AND IMPLEMENTATION

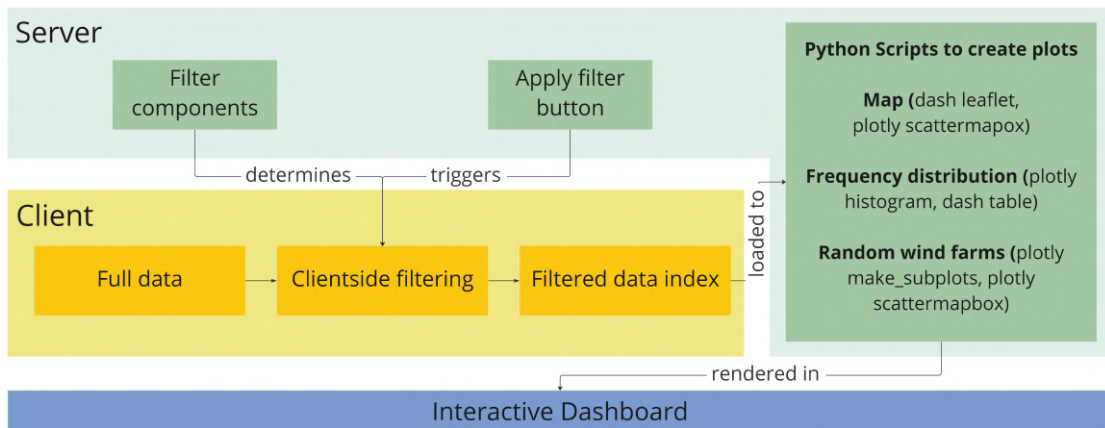


Figure 6.9: Simplified prototype architecture

The switch to `clientside-callbacks` reduced the time for filtering from an average of about 55 seconds to about 3 seconds. However, the initial loading of the data in the browser caused a delay, as all 43 MB had to be uploaded. However, the longer time for the initial loading was taken into account to provide a better user experience later.

A more detailed overview of the Dash application structure is shown in Figure 6.10. The round components of the graphic are Dash components with their id in the middle. The square elements are the callback functions, with their response time in the middle. The circles surrounding the elements outlined in red refer to clientside elements, and the green ones are on the server side.

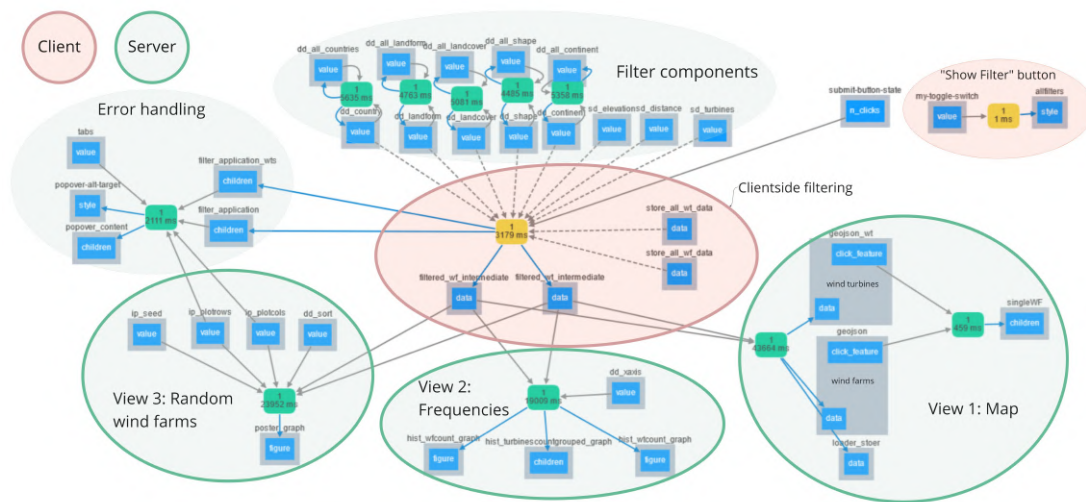


Figure 6.10: Detailed prototype architecture



### 6.3.3 Views Implementation

Each view has its server-side callback function, which inputs the filtered data indexes from the `dcc.Store`. Depending on the view, they have additional input variables.

#### Map View

The map for the overview was created using the `Dash Leaflet` library. The marker clustering was done using the library `supercluster`, a fast geospatial point clustering library [81]. The map callback function takes only the filtered data indexes as input, selects the correct records from the dataset and returns the variables `WFid`, `long`, and `lat` in a `Geobuf` data structure. `Geobuf` is a compact binary encoding for `GeoJSON` data and saves much memory. The `Geobuf` data is passed directly to the `GeoJSON` dash leaflet component in the output, which is embedded in the map.

An additional callback is created to plot the single wind farms located on the map's right side. Its callback is when a wind farm marker is clicked on the leaflet map. The callback input is the `WFid` of the selected marker, upon which a `scattermapbox` of the selected wind farm is created. The `scattermapbox` plot corresponds to the individual wind farm plots from the Random wind farms view and uses the same visual encoding.

#### Frequency distributions

The second view has a simple composition. The histograms are created using the `Plotly` function to create histograms. The table is drawn from the summary statistics about the wind farm sizes and uses the `Dash DataTable` [82] for rendering. Its callback has the additional input of the variable according to which the frequencies are displayed. Unlike the filter functions, the visualizations change every time a change is made to this variable.

#### Random Wind Farms

The last view uses multiple sources to create the visualization and takes the most callback inputs. The inputs are the variable by which the illustrated single wind farms should be sorted, the seed to draw a random sample and the number of rows and columns of the raster.

The raster-shaped structure of the plot is provided by the `Plotly` option to create subplots with `make_subplots`. For this, the number of columns and rows is specified with a default maximum of ten columns and ten rows. The individual wind farm plots are created equal to those of view 1 using the `Plotly` function `scattermapbox`. Random wind farms are drawn in advance, with the seed from all the wind farms matching the filter settings. On the `scattermapbox` plots, the turbines belonging to the selected wind farms are marked in red, and all remaining surrounding wind turbines with a radius of 1° latitude and longitude are marked in yellow.

## 6. PROTOTYPE DESIGN AND IMPLEMENTATION

---

The map layer used was designed using `Mapbox Studio` [83]. The grid overlaying the wind farms was created with `Turf.js`. The grid centre corresponds to the centre of the wind farm, and the size of the grid covers all the turbines included in the wind farm. The lines of the small squares in the grid correspond to a constant distance of one kilometre, so each square of the grid corresponds to one square kilometre of area.

# Evaluation

The evaluation took place as a user study involving ten participants. Users were given several assignments during one-on-one online interviews, to which they responded by speaking the answer aloud. While interacting with the tool, they were also encouraged to think aloud. Different evaluation techniques were employed, both qualitative and quantitative methods:

## 1. Quantitative:

- Accuracy in task-solving [84]
- Feedback in form of Questionnaire (ICE-T heuristic) [33]

## 2. Qualitative:

- Open questions for insights generation and feature detection [85]
- Feedback in verbal form

The feedback Sections were important for assessing the overall performance of the developed prototype. The task accuracy and open question assignments were necessary for answering the research questions.

In this chapter, we will present the users we interviewed, the different assignments, the general structure of the interviews and the outcomes.

## 7.1 Users

Interviews were conducted with ten participants. Of these, three persons were domain experts, i.e. in the field of wind energy, six persons were visual analytics experts and one was an expert in the field of Geography. The assignments to solve were the same for all

participants, except that domain experts got additional feedback questions regarding the tool's usability in their work.

### **Domain Experts**

The domain experts were referred by the expert who supervised this project, Dr Johannes Schreiber. They are all actively working in the field of wind energy and have a related PhD. They stood as crucial participants in putting our enriched data and visual system to trial since they have comprehensive knowledge about the domain [33]. Through them, the used characteristics describing wind farms could be validated, and additional ones determined. Furthermore, they could also confirm if the end-users are helped with the solution by doing tasks they could not do before.

### **Visualization Experts**

Dr Victor Schetinger mediated the visualization experts. They were all former or active researchers in the Centre for Visual Analytics Science and Technology (CFAST) which is part of the Vienna University of Technology (TU Wien). Tory and Möller [86] found that visualization experts can provide valuable feedback on visualization designs and methods, which is why they played an essential role in our evaluation. Their main function was to evaluate the overall design, functionalities, user interface and latency of the tool.

### **Geography Expert**

Another participant was a researcher in geography and regional research with expertise in working with geographical data and developing GeoJSON applications. This person was sought as a participant to provide valuable feedback from a geographic perspective and to discuss potential enhancements to the mapping and map projections of the tool.

## **7.2 Interview Structure**

The interviews took place as online meetings in MS Teams and lasted about one hour each. They involved the visualization researcher and one participant per session. The meetings consisted of the participants solving assignments to which they responded using the "think-aloud" method. With this method, they are given a task/question and then instructed to speak aloud the answer and their thoughts on the way there. To capture spoken dialogue, we recorded the sound and computer monitor while they were sharing their screen while using the prototype. For this, we asked them if they consented and explained that we would delete the recordings after the evaluation.

The visualization researcher acted as a guide for the visualization, i.e. if the participant needed help and explicitly requested it, the visualization researcher assisted in solving the assignment; otherwise, the researcher's role was to provide initial explanations about the data and tool.

The interview structure and the duration of the different steps are listed below.

**Step 1:** Contact participant and send manual of prototype (before the interview)

**Step 2:** Introduction (15 min.)

- General introduction to the topic
- Presentation of prototype and data source
- Demonstration on how to answer three example tasks

**Step 3:** Free usage of tool (5 min.)

**Step 4:** Task accuracy assignments (20 min.)

- P2: characteristics and relationships
- P3: comparison of wind farm topology

**Step 5:** Open questions (10 min.)

- P2: Discover relevant features and insights

**Step 6:** Open feedback questions (10 min.)

**Step 7:** ICE-T Questionnaire (after the interview)

The participants were not familiar with the prototype before the interview but were sent a manual they could study beforehand. The manual describes the data source and variables, including their categories. In addition, it explains the prototype's main functionalities and shows how one can use it to solve three example tasks. The manual is displayed in Figure ??.

After the participants confirmed that they would participate in an interview, we scheduled an appointment and sent them the prototype manual via mail. During the interview, the participants were first given a short introduction and an outline on what to expect. The prototype and data were presented, and the participants were instructed on how to answer the assignments. They were also shown how to use the prototype to answer questions by solving three example tasks similar to the ones they would face later. At this point, they were also asked for permission to record.

Then the actual key phases for the interview started. The participants were sent the link to the tool and initially given five minutes to get acquainted with the it, during which they could already use the visualization to gain insights. The most extended phase came in the beginning, namely answering the tasks for the problems from P2 and P3. Two open-ended questions followed to allow the participants more opportunities for open-ended exploration to gain insights and thus identify characteristics of the problem

from P1. Finally, the participants entered the feedback phase, for which they no longer had to use the tool but rather engaged in a conversation with the researcher.

After the interview, the respondents received an online questionnaire by email, which they were asked to complete. The questionnaire was developed from the ICE-T methodology and served to obtain a quantitative measure of the quality of visualizations. The assignments are clarified in the following Section.

### 7.3 Assignments & Results

As discussed in the introduction of this chapter, the assignments served different purposes. On the one hand, the aim was to solve the research questions by detecting features, generating knowledge about wind farms (RQ1), and accurately solving tasks (RQ2, RQ3). On the other hand, the assignments should evaluate the tool's performance through verbal and questionnaire feedback.

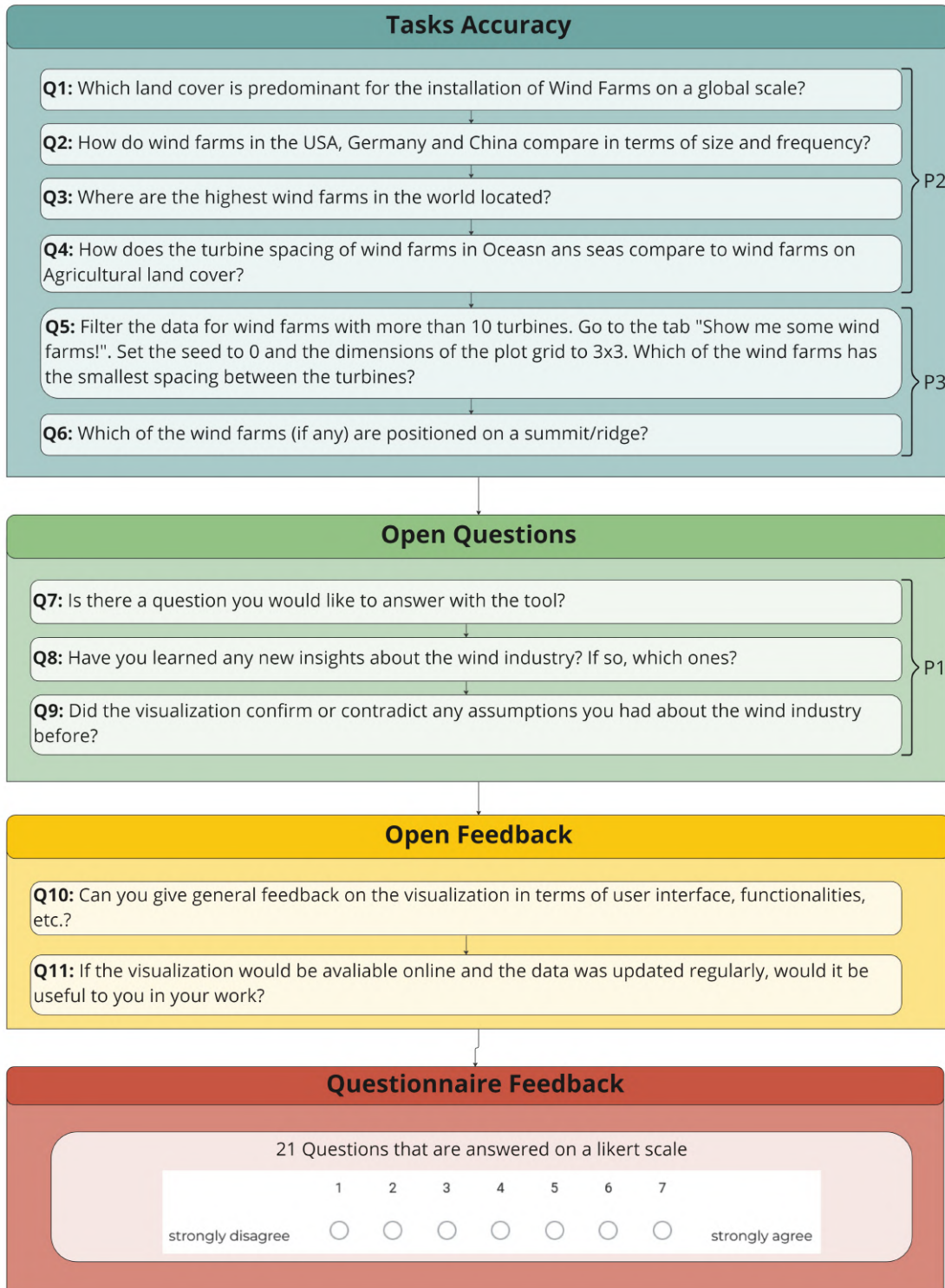


Figure 7.1: Overview of tasks and interview questions in user study evaluation

### 7.3.1 Tasks

The tasks were posed to the participants as questions to which they had to find an answer. They aimed to analyse the data properties and relationships and characterizing individual wind farm topology with the corresponding problem domains P2 and P3.

How accurately the tasks were solved was measured by how many participants answered the questions correctly. An answer is considered correct if the participant understands how to answer the question without help and arrives at a correct solution.

#### **P2: Data characteristics and relationships**

The tasks from P2 were designed to test whether the participants could use our visualization system to assess the validity of the claims from the literature in Chapter 2 by using the EDWin dataset. Therefore, we developed the tasks to match the claims, which led to four questions.

- Q1 Which *land cover* is predominant for the installation of wind farms on a global scale?
- Q2 How do wind farms in the USA, Germany, and China compare in terms of size and frequency?
- Q3 Where are the highest wind farms in the world located?
- Q4 How does the turbine spacing of wind farms in Oceans and seas compare to wind farms on Agricultural *land cover*?

#### **P3: individual wind farm topology**

The questions for the domain problem P3 were specific to the Random Wind Farms view and were based on a sample of nine farms equal for all participants. Identical sampling was achieved by applying the same filter and using the same seed, as was defined in the question. The wind farm sample is displayed in Figure 7.2. Based on this grid of farms, participants were asked to determine the wind farms with the following characteristics:

- Q5 Filter the data for wind farms with more than 10 turbines. Go to the tab "Show me some wind farms!". Set the seed to 0 and the dimensions of the plot grid to 3x3. Which of the wind farms has the smallest spacing between the turbines?
- Q6 (Refers to Q5) Which wind farms (if any) are positioned on a summit or ridge?



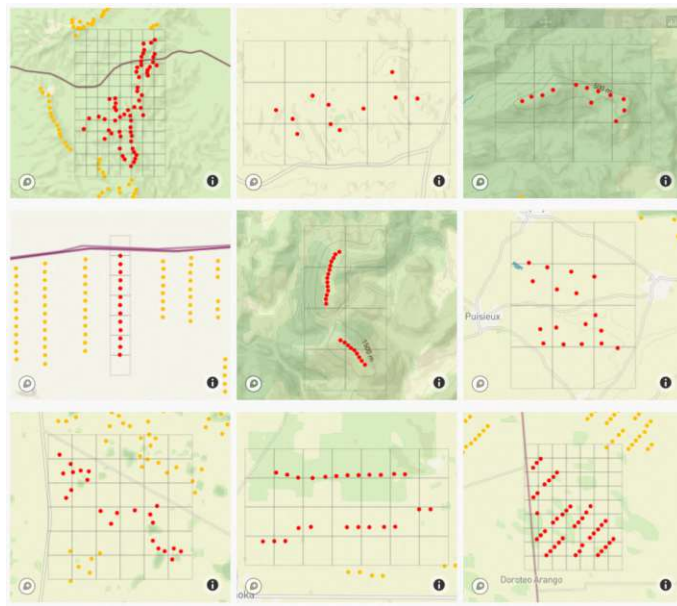


Figure 7.2: Random sample of wind farm configurations for Q5 and Q6

## Results

**Q1** and **Q3** were straightforward and answered correctly by all participants without help. More difficulties were encountered with **Q2** and **Q4**, where comparisons between several variable factors had to be analyzed. Intuitively, most participants assumed that if they filter for factors of a variable, the frequency view would group them accordingly for other variables, but this was not the case. **Q5** and **Q6** were answered correctly by most participants, although we received many comments on the interpretability and scaling of the graphs for these questions. Furthermore, instead of only answering the questions by visual inspection, some participants used the hover information to help them.

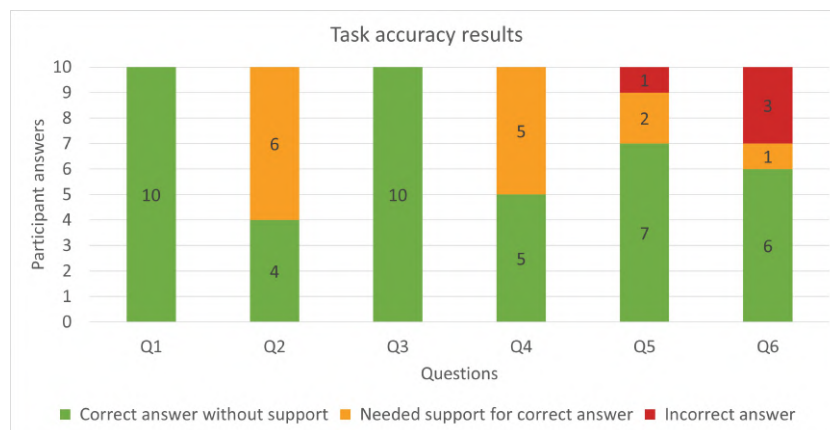


Figure 7.3: Task accuracy results of user responses

**Q2** was only solved by four of the ten participants without help, which was surprising since the solution could be read directly from the frequency bar chart of the three countries and the adjacent table. Most participants found the number of wind farm comparisons but struggled with the size comparison. They intuitively put the variable "number of turbines" on the y-axis, giving them the joint distribution of the wind farm sizes of the three countries.

At this point, these six participants asked for help or if they were doing something wrong. They did not seem to understand the purpose of the table right away. After being reminded that the table refers to the wind farm sizes of the factors, all participants understood and could answer the question without any problems.

**Q4** presented similar problems, but here we were expecting them more. Five participants needed help to find the correct solution. The approach of most participants was correct at first, the frequency table with turbine spacing on the y-axis. However, initially, most participants looked at the joint distribution of "Agricultural" and "Oceans and Seas" *land cover*, expecting them to be grouped automatically. Half of the participants realized they would have to look at these individually. The rest asked for support at this point.

An interesting observation we found was that one of the participants tried to initially answer **Q2** and **Q4** questions with the map view. The person was successful to discover number of wind farms according to the countries, but had to change his strategy to compare the wind farm sizes and the turbine spacing.

**Q5** was successfully solved by seven of the participants, five of whom found the solution directly after a quick overview of all wind farms. The rest were initially confused by the scaling of the plots and found it difficult to solve the tasks by purely visual means.

For **Q6**, there were a total of 3 wind farms located on a peak/ridge, with six people finding them all. One person found them only with the help of the interviewer, and the rest found only some of the wind farms or gave incorrect answers. Four people made use of the hover information and three people sorted the wind farms according to the land form.

### 7.3.2 Open Questions for feature detection and insights generation

The open questions allowed the participants to freely use the tool to generate their insights and to have domain experts speak about what features can be used to characterize wind farms, what insights they were able to find and which they would be interested in finding.

Before starting the assignments, the participants were already given 5 minutes to work freely with the tool. Then, after solving the tasks section, we asked an open-ended question:

Q7 Is there a question you would like to answer with the tool?

With this question, the participants could explore all questions of interest to gain more insights. Domain experts were asked what additional features they would be interested in and what features they consider important when characterizing wind farms.

Afterwards, we asked them the following questions:

Q8 Have you learned any new insights about the wind industry? If so, which ones?

Q9 Did the visualization confirm or contradict any assumptions you had about the wind industry before?

## Results

The responses from the domain experts and other participants were evaluated separately. In the case of the non-domain experts, we were interested in what insights could be gained by people who have little or no knowledge of the present situation of wind installations. The domain experts' responses were intended to validate the insights gained and to show whether they, too, were able to gain new insights despite their substantial knowledge of the subject. In addition, an in-depth qualitative discussion was held with the domain experts. They were asked to argue which of the existing, but also which additional variables they consider relevant to characterize a wind farm and which local and global patterns they depict according to their knowledge and following the visualization results.

**Non-domain Experts:** For **Q7**, participants repeatedly investigated how wind farms are distributed across specific continents and countries, focusing on the participants' places of origin or their current residences. Other interests included the distribution of wind farms across land areas, e.g., oceans and seas, their distribution across specific countries, and the distribution of stand-alone turbines. One person, for example, was in South America and was, therefore, more interested in the distribution of wind farms across countries in South America and over the different *land covers*.

An interesting question we found was about the location of specific wind farms, which were often discussed during the election campaign that was taking place in a given country at the time of the interview. The respondent said they had heard about these wind farms several times but did not know where or how big they were.

In the discussion that followed about whether insights were discovered for **Q8**, all non-domain experts answered yes. They knew little or nothing about the wind industry, so they felt that any answer to the tasks would yield insights. The comprehensive insights gained by the non-domain experts are listed below:

- USA and China are leading powers in the number wind installation
- There are far fewer wind installations in the southern hemisphere than the northern hemisphere
- There are many more wind installations around the world than thought

- There are large wind farms in the north of Brazil
- Brazil is the leading wind energy producer in South America
- There is a variety of shapes that wind farms can have, not only arrays

**Domain Experts:** The domain experts had a different and more explicit approach to answering Q7 and Q8. They had concrete questions they wanted to answer with the tool since they have more knowledge about the domain, and they saw this as a chance to answer some of them. The questions they wanted to answer were organized as follows:

- The distribution of wind farms over complex terrain
  - particular interest in high elevations, ridges, cliffs, and valleys.
  - how many wind turbines are located on complex terrain in which countries (experiments with many *land covers* and *landforms*)
- Where turbines are located on simple flat terrain (particularly in Germany)
  - flat terrain wind farms can be used to install wind-liDAR systems (works better due to more even wind flows)
- Detection of wind farms in places where they were not known to exist
  - Antarctica, American Samoa, Vanuatu
- Analysis of wind farms in South Africa:
  - where wind farms are located, on the coast or in the country's interior
  - how many are in wind turbines are in farms and how many are stand-alone
  - over which *land covers* they are distributed
  - which shapes of wind farms are most represented
  - the distribution of the sizes of the wind farms
  - where the biggest wind farm is
- Analysis of global hot spots of wind farms:
  - search for hot spots in the USA
  - search for wind farms in California, which are known to be the largest
  - analyzes hot spots and cold spots by different *land covers*
- Search for trends of *land covers* in different countries
- How many wind turbines are installed in Germany, and where they are located

- Search for wind farm projects they were involved in

One can see that the questions asked by the domain experts were much more strategic and purposeful than those asked by the non-experts. All three experts mentioned that seeing such extensive data with the chosen variables was unprecedented. There was particular interest regarding the variables *land cover* and *landform*. One of the experts commented: "*yes, that it is possible to quantify the distribution of wind turbines by landform and land cover is really useful*". The fact that they had never seen such information in conjunction with wind infrastructure data may be responsible for many questions revolving around them.

When asked about insights they had gained, the domain experts answered the question based on the answers to the questions above. So, for example, they now know there are wind farms in Vanuatu or how wind farms are distributed on complex terrains in Germany, which are very concrete examples, so we will not list them all. One domain expert, described his experience in a nutshell: "*I had new insights in the sense that all the world data is there, and that is quite powerful.*"

Overall, valuable insights from the domain experts were the following:

- India has many very large wind farms
- Most wind farms are located on agricultural land
- Wind farms in Germany are relatively small, compared to other countries
- Flat terrain is the most represented *landform* for wind turbines

Many of these insights were familiar to the domain experts, with the tool now allowing them to confirm them: "*I knew instinctively that wind farms in China tended to be larger than in Germany; I think actually to see numbers about that is really interesting*". However, they also learned things that were not expected: "*I knew that a lot of turbines were on flat terrain, but I didn't think that so many were identified being on other types of terrains, that actually quite surprised me. It looks in total less than than half of the turbines are on flat terrain*".

All three found the included variables very informative, but mainly for general purposes. Any technical information about the farm's turbines, as well as information about the wind flows would increase the value of the visualization even further. The following features emerged from the interviews as additional characteristics of wind farms.

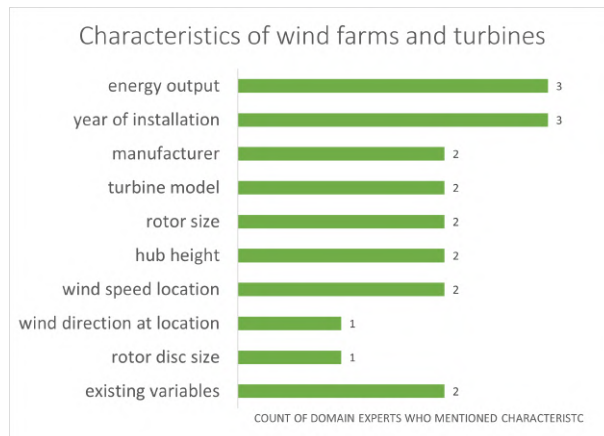


Figure 7.4: Frequency of important Wind Farm and Turbine characteristics determined by Domain Experts

It was clear to the experts that it would be challenging to obtain most of these data on a global scale, and that it would require a considerable investment in time and resources. Nevertheless, they recommended databases from which some data could be extracted. Further, one expert mentioned that the databases could be used to estimate some of the unknown characteristics, such as power capacity. All experts view power capacity as an essential characteristic of a wind farm, so its estimation could be a challenge to inspire future studies.

### 7.3.3 Feedback

#### Verbal feedback

After completing the tasks and open questions, we asked participants for feedback regarding the tool's utility and user interface/functionality. Domain experts were asked an additional question about the tool's usefulness for their work.

Q9 Can you give general feedback on the visualization in terms of user interface, functionalities, etc.?

Q10 (Domain Experts only) If the visualization data was updated regularly, would it be useful to you in your work?

#### Results

Participants gave a lot of feedback about the user interface (UI) and functionalities, the most common points for improvement are shown in Figure 7.1.

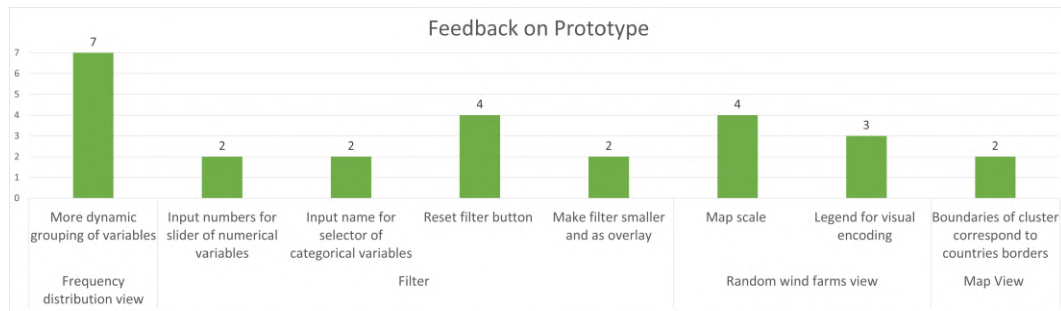


Table 7.1: Frequency of Prototype feedback from users on potential improvements and additions

Given that many participants had difficulties with **Q2** and **Q4**, and the lack of grouping by an additional variable in the frequency distribution view triggered this, it was most often mentioned as a suggested improvement for the prototype. We acknowledge this as a shortcoming of the prototype since it is not practical to remember previous results when performing tasks.

Some participants had suggestions for improving the filtering interface. They were displeased by its size, given that the visualization did not fit on the website when expanded. Thus, they suggested making it smaller or as an overlay. Also, some people missed a reset filter button, and others thought that the apply filter button would be better placed at the end of the filters. Finally, for the numerical sliders and the categorical selector, the possibility to input the values using the keyboard would be beneficial and save some time.

The random wind farms view was also often the subject of suggestions for improvement, primarily related to its interpretability. Thus, the participants suggested either the same scale or an indication of the scale on the wind farm maps. Further, a legend describing the visual encoding would be helpful.

On two occasions, participants suggested that the clustering of the map view should be adapted to the country boundaries, i.e. that country-specific clusters should be created. Other ideas to extend the visualization which were mentioned only once are the following:

- percentual information for the frequency distribution
- interlinking from map view to google earth or similar websites
- the possibility to give feedback about the data, i.e. report if a turbine is classified on a wrong *land cover*
- the integration of the frequency view and the map view to one common view
- enabling the search for specific locations



- allow the selection of wind farms in the frequency distributions and display them on the map view

Question **Q10** was only addressed to domain experts. They all answered affirmatively; they could imagine that the visualization could be helpful in their work. They also mentioned areas in which they thought the visualization could be helpful or mentioned functionalities that would make it useful for specific areas.

To quote one of the experts' direct answer: "*I could see this being really useful for doing the introduction to wind energy courses at university, I can see it being really interesting for people in national wind energy organizations or governments who want a standard database they can use with relatively easy access, I can see it being really useful for start-ups, either completely new businesses or people coming in from other businesses, and also for consultants in wind energy. [...] Everything is going to go digital, and it is going to save the world, and few people understand what effective access to data may actually allow them to do; this kind of visualization is actually a really nice example of that, and it also lets you discover some interesting things about the wind energy industry, like where are turbines, what are the challenges that people might face.*"

One domain expert searched very often for wind turbines in google maps or google earth, and therefore emphasized that it would be helpful for him to link the turbine coordinates with google earth to view them there directly.

Another domain expert is a lecturer at the university and said that the visualization would be beneficial as an introduction to the topic for his students and for the students to be able to do projects with it. In total, two of the experts mentioned that it would be convenient in a lecturing context.

One of the experts submitted a paper a few weeks before the interview on "wind farming on complex terrain". He said he would have needed this exact data about the frequency distribution of turbines over *elevation* and "cliffs", "valley" and "ridge" *land forms* in various countries. He was considering a similar way to obtain such information to include in the study but realized that it would have taken him at least two weeks of effort, so he decided against it. However, due to this, he was familiar with the data sources we used in the study and was very impressed. Furthermore, he liked that the plots could be downloaded as images and even asked if he could use the visualization for private use.

Nevertheless, the experts also mentioned that the tool is very exploratory with a general purpose. In order to use it in a way that makes the work more efficient and accessible, it would have to be conceived in such a way that it serves specific customers or markets to achieve specific goals. These goals vary greatly depending on the customer's background, and their exact questions need to be carefully analyzed and understood. For example, specific purposes would be policy-making for wind infrastructure in regions or countries or wind energy developers looking for expansion opportunities or turbines that need to be replaced.



## Questionnaire

To obtain a value for the utility of our visualization in addition to task accuracy and insights-generation, we adopted the ICE-T methodology from the paper "A Heuristic Approach to Value-Driven Evaluation of Visualizations" by Emily Wall *et al.* [33] to design a questionnaire. It evaluates a visualization in terms of quality based on four components: insight, time, essence, and confidence. For each component, core concepts are defined, which the authors refer to as guidelines and which they further break down into heuristics. The heuristics are designed to reflect the components. They are ultimately the underlying questions of the questionnaire. The heuristics are rated on a Likert scale from 1 to 7, with 1 standing for "strongly disagree" and 7 for "strongly agree".

## Results

The questions and results of the questionnaire are shown in Figure 7.3. The responses of all participants were averaged for each heuristic. We calculated the average of all heuristics to obtain a total cumulative visualization score. In Figure 7.2, the scores are aggregated for the guidelines and the components. The authors of the heuristics-based questionnaire say that good visualizations should have a cumulative average score of 5 or more. Weak visualizations have a score of 4 or less. Everything in between is considered neutral. Specific weaknesses are further indicated by the fact that individual heuristics and guidelines score less than 4.

Our visualization received an overall score of 5.93, which corresponds to the authors' requirements for a good visualization. The insight and time components received the best scores, each higher than 6. On the guideline "The visualization provides mechanisms for quickly seeking specific information", we received an exceptionally high score of 6.15. This indicates that our visualization is easy to understand and compactly designed to provide information quickly, which was also one of our goals. The only heuristic that received a score lower than 4 was that of the guideline "The visualization helps understand data quality" or the specific heuristic "If there were data issues like unexpected, duplicate, missing, or invalid data, the visualization would highlight those issues". This indicates that our visualization had the weakness of not clearly evidencing missing values and data quality.

## 7. EVALUATION

---

components	guidelines	guidelines scores	components scores
Insights	The visualization facilitates answering questions about the data	5.96	6.06
	The visualization provides a new or better understanding of the data	6.11	
	The visualization provides opportunities for serendipitous discoveries	6.11	
Time	The visualization affords rapid parallel comprehension for efficient browsing	6.00	6.09
	The visualization provides mechanisms for quickly seeking specific information	6.15	
Essence	The visualization provides a big picture perspective of the data	6.06	5.94
	The visualization provides an understanding of the data beyond individual data cases	5.83	
Confidence	The visualization helps avoid making incorrect inferences	6.11	5.49
	The visualization facilitates learning more broadly about the domain of the data	6.11	
	The visualization helps understand data quality	3.63	

Table 7.2: Guidelines and components scores

components	guidelines	heuristics	$\mu$	$\sigma$
Insights	The visualization facilitates answering questions about the data	The visualization exposes individual data cases and their attributes	6.22	0.79
		The visualization facilitates perceiving relationships in the data like patterns & distributions of the variables	6.11	0.57
		The visualization promotes exploring relationships (between individual data cases as well as different groupings of data cases)	5.56	1.07
		The visualization helps generate data-driven questions	6.11	0.57
		The visualization provides a new or better understanding of the data	6.11	0.57
		The visualization helps identify unusual or unexpected, yet valid, data characteristics or values	6.11	0.57
Time		The visualization provides useful interactive capabilities to help investigate the data in multiple ways	6.11	0.57
		The visualization provides opportunities for serendipitous discoveries	6.44	0.68
		The visualization uses an effective representation of the data that shows related and partially related data cases	5.78	1.23
		The visualization provides a meaningful spatial organization of the data	6.22	0.92
		The visualization affords rapid parallel comprehension for efficient browsing	5.78	0.79
		The visualization shows key characteristics of the data at a glance	6.22	0.42
Essence		The interface supports using different attributes of the data to reorganize the visualization's appearance	5.56	1.07
		The visualization provides mechanisms for quickly seeking specific information	6.67	0.47
		The visualization supports smooth transitions between different levels of detail in viewing the data	5.89	0.87
		The visualization avoids complex commands and textual queries by providing direct interaction with the data representation	6.22	0.79
		The visualization provides a big picture perspective of the data	6.22	0.63
		The visualization presents the data by providing a meaningful visual schema	5.44	1.07
Confidence		The visualization facilitates generalizations and extrapolations of patterns and conclusion	6.11	0.99
		The visualization helps understand how variables relate in order to accomplish different analytical tasks	6.11	0.99
		The visualization provides an understanding of the data beyond individual data cases	6.11	0.87
		The visualization helps avoid making incorrect inferences	3.63	1.73
		The visualization promotes learning more broadly about the domain of the data	5.93	
		The visualization helps understand data quality		
<b>Cumulative Visualization score</b>			<b>5.93</b>	

Table 7.3: ICE-T Methodology questionnaire with results

## 7.4 Summary & Discussion

One of the most exciting aspects we found was seeing how diverse visualization experts, domain experts, and geography experts used the tool to gather information and what kind of feedback they gave on the functionalities. Of course, domain experts had much more prior knowledge and assumptions about the data, which they could now prove. Therefore, they went on direct solution finding and answered questions in a very structured way towards a goal. They knew what to do with the tool right from the start and naturally suggested fields of application in which the tool could be integrated. The assessment of the tool's functionalities was more directed toward what could be added to make it more beneficial for themselves, i.e. adding interlinking to google earth.

On the other hand, the visualization experts were much more concerned with the user interface and the visualization design, so their feedback focused on these issues. The questions and insights they would seek were more generic than those of the domain experts; in a way, the first thing that came to their mind was that almost everyone went directly to their home countries to see how wind farms were distributed there. In addition, they were much more concerned with playing around with the functionalities to see what was possible rather than getting factual information out of the data.

The geography expert was very interested in the tool. He did not know much about turbines but found analyzing them in terms of *land covers*, *landforms* and *elevations* very intriguing. He was particularly interested in how the data was obtained and what projections were used. After the interview, he even asked to be updated on the project's progress.

The diversity of backgrounds made it all the more interesting to involve them all in answering the questionnaire, as the quality of the tool can bring benefits to be assessed from multiple perspectives, the absence of which the authors in [33] considered a limitation in their evaluation.

Overall, the visualization produced satisfactory results in finding solutions to the assignments we gave them. The strength of the visualization is that it shows the relationships and information at a glance and has an easy-to-use interface. Moreover, the completeness of the data and the variety of features makes it useful for generating insights and questions from the data. The main shortcomings we identified were the lack of dynamic grouping for the frequency view and the lack of legend and scaling in the random wind farms view.

# Limitations & Future Work

Based on limitations we encountered during the project, the thesis has identified several opportunities for future research.

**Data enrichment.** Interviews with domain experts have shown that there is still room for improvement in adding variables to characterize wind farms. In particular, variables for which global datasets are available, such as historical average wind speed and current wind speed at the site. Adding these variables to the visualization as overlays would significantly enhance the insights and decision-making capabilities for domain experts, as they would provide them with a more comprehensive understanding of the wind conditions at the site. This information is important for characterizing wind farms because it can be used to predict the potential energy output of the farm, optimize the design and placement of wind turbines, and detect new installation areas.

**Data quality.** The data obtained by enrichment can be affected by errors. These can be related to the turbines themselves, i.e., whether they are still up to date, or to the information about the turbines. In the latter case, turbines are mainly affected if they are located in border areas between two factors ( e.g., turbines located on the coast are located at the border between oceans and seas *land cover* and a different one). These are misclassified due to the degree of resolution of the dataset or by changing landscape conditions, e.g., the increase of the sea level changes the land area. Therefore, feedback on turbine information in the visualization itself, as proposed by a domain expert, would be highly beneficial. It would enable a more effective data cleaning and improve the overall quality of the dataset in all respects: Factual existence of wind turbines, correct clustering in wind farms, correct classification of *land cover* , land shape, shape, country, and continent. It would be an valuable step towards the construction of an official global database on the global wind infrastructure.

**Estimation of rotor size and power output.** During our study, we encountered an insight that could be used to estimate turbine rotor size and power output. For the data that have complete information on power output (124,059 turbines) and rotor size (29,663 turbines), we plotted the variable turbine spacing against each, as shown in Figure 8.1 and Figure 8.2. The visualizations suggest that there may be relationships between turbine spacing and rotor size, thus, turbine spacing and power production. We believe that, along with data on wind speed at the site, this may allow estimates of energy production for wind farms worldwide.

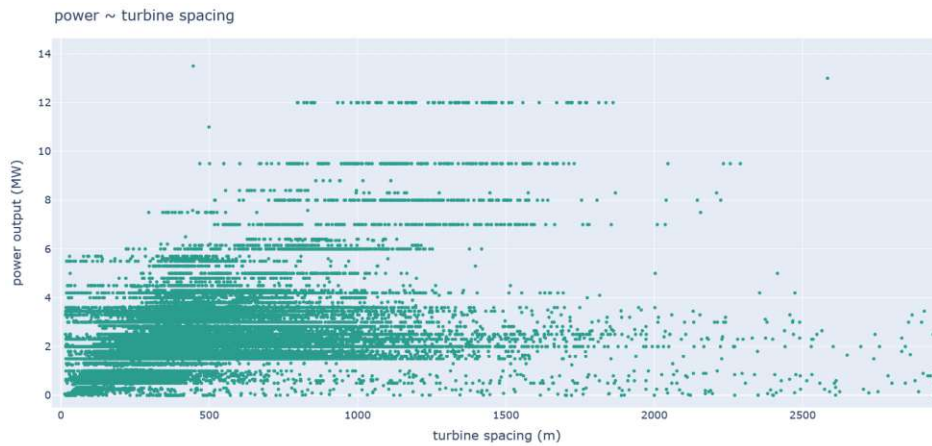


Figure 8.1: Scatter plot of turbine spacing vs. power output

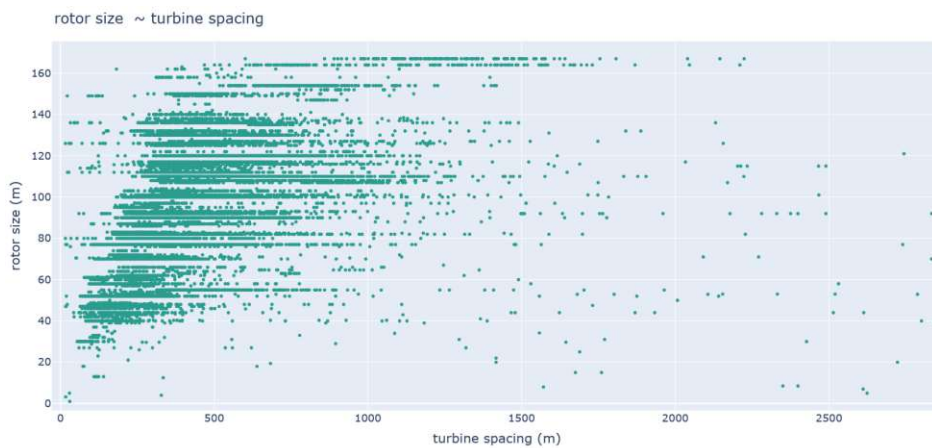


Figure 8.2: Scatter plot of turbine spacing vs. rotor size

In order to investigate the relationship between power output or rotor size and turbine spacing, we conduct correlation analyses using Pearson's product-moment correlation coefficient. The results of these tests reveal a statistically significant correlation at a 95% confidence level, as indicated by p-values less than  $2.2e-16$ .

---

The correlation coefficient ( $\text{cor}$ ) for the power output and turbine spacing test is 0.1921765, indicating a weak positive correlation. Similarly, the correlation coefficient for the rotor size and turbine spacing test is 0.3604944, which suggests a moderate positive correlation between the two variables. We also performed a linear model analysis to assess the ability of rotor size and turbine spacing to predict power output. The results show that both rotor size and turbine spacing have a strong positive correlation with power output and are statistically significant in explaining the variation in power output. As rotor size and turbine spacing increase, power output is likely to increase as well.

**OpenStreetMap data.** The analyzed tags and other ones, which we did not include in the analysis, could be adjusted and included in the visualization. Despite significant incompleteness, we still found relatively complete data in some countries, mainly where OpenStreetMap is better known. Furthermore, information could be retrieved from the OSM History Viewer [87]. It stores when and how nodes were changed, moved, created, or removed, and all tags that were removed or added to elements. This data could provide insights into the temporal evolution of global wind turbine infrastructure.

**Wind farm clustering.** Defining wind farms based on spatial characteristics and to identify them using clustering algorithms deserves deeper investigation. Dunnett's approach in [7] in determining parameters for clustering wind farms is an appropriate course of action. However, it requires improved data and an acknowledgement that wind farms can be very different depending on variables such as *land cover*, *landform*, and *country*. For example, offshore wind farms are very different from other wind farms because of the large distances between turbines, so different parameters are needed to cluster them. To this end, spatial analysis on more representative datasets of known wind farm projects which consider factors related to land characteristics would be required.

**Clustering of shapes.** The clustering of shapes did not work as well as we had hoped, so we left out the results to evaluate the prototype. During the development of the model, we found that mainly wind farms with similar sizes were grouped, i.e. the number of wind turbines was more decisive than the actual shape. We obtained acceptable results only for simple shapes such as lines, although not all line-shaped wind farms were detected. An unsupervised clustering approach without knowledge of the type and number of groups is challenging to evaluate and refine. Therefore, applying a model to data with predefined shapes of wind farms would be more prudent. That would require working with experts to manually classify and name a large enough number of wind farms so that the model knows what to look for.

**Selection of implementation environment.** After the prototype development, we would switch to another, more customizable implementation environment since we encountered some limitations using Dash. Among them is the grid that overlays the wind farms in the individual wind farm encoding plot. The initial plan was to



overlay the map with a coordinate system, but Dash's functionalities did not enable that. Further, the wind farms in the map view should be clustered as polygons so that wind farms and turbines could be viewed simultaneously, but Dash's built-in functionality also hindered this.

**Revision of the visualization and publication for general use.** The visualization still needs modifications in several ways, as described in Section 7.4. One key point is that the frequency view of our prototype lacks a more flexible grouping behaviour for the histograms and bar charts. Including such would allow more in-depth analysis and thus should be kept in mind for the further development of the visualization. In addition, the random wind farms view also requires significant adjustments. Eventually, the visualization and the data should be made public for general use. Therefore the data must be published according to the FAIR principles [88] and the visualization interface needs additional information about data sources, licensing, contact to the authors, and an embedded manual on how to use the tool.

**Improving Wind Farm Visualization Encoding** The Random wind farms view described in Section 6.2 has effectively displayed wind farms and their installation area and layout; however, some refinements could be made to display the intended elements more effectively. Further research would be needed to make these improvements. As already mentioned, improvements in the grid encoding would enhance its functionality. By transforming the grid into a coordinate system, it would become more flexible and easily understandable for users, allowing them to quickly access distances and other information. Additionally, more research could be done to explore other ways of visualizing wind farm data, which could lead to the development of a widely accepted and adopted standard for displaying it in the industry.



# Conclusion

## 9.1 Answers to the Research Questions

We collaborated with a domain expert and a visualization expert to create an enhanced dataset of wind farms and turbines, as well as an interface for visualizing the data on both a global and wind farm-specific level. We conducted a user study, testing the prototype and gathering data through qualitative interviews with wind energy experts and measuring task-solving accuracy with visualization, geography, and domain expertise experts. We evaluated the results and found that the solution could effectively perform the tasks for which it was designed. Through in-depth discussions with domain experts, we also gained valuable insights into the needs and priorities of the wind energy community for future global development. This methodology allowed us to gather valuable data to help us answer our research questions.

**RQ1** What features can be used to characterize wind farms across the world, and what local patterns do they present?

Conversations with domain experts revealed that features characterizing wind farms and turbines consist of technical, temporal, terrain, and weather data.

**technical:** Energy output, Manufacturer, Turbine model, Farm layout, Turbine spacing/density, Rotor size, Hub height, Rotor disc size;

**temporal:** Year of installation;

**terrain:** Land cover, Landform, Elevation level, Country/Continent;

**weather:** Wind speed, Wind direction;

Our data enrichment focused on terrain data; thus, the other areas would be strongly complementary. Yet, for technical data, in particular, there still needs to be more possibilities to include them in the data in a more comprehensive way.

Global and local patterns are manifold, and many still need to be discovered. Yet the participants found and confirmed some patterns based on the included characteristics that are worth mentioning:

- The US and China are leading powers in the number of wind installations.
- Germany has much smaller wind farms than the US and China.
- In the southern hemisphere, there are much fewer wind installations than in the northern hemisphere.
- India has many huge wind farms.
- The highest wind farms are located in China.
- Flat terrain is the most common *landform* for wind installations.
- The predominant *land cover* for wind installations is agricultural land.
- Wind farms on the ocean have a greater distance between turbines than wind farms on land.

**RQ2** What visualization technique is appropriate to explore the claims about wind farms found in the literature using the EDWin?

The user study participants successfully solved the tasks to check the validity of the posed claims about wind farms in Chapter 2, even though some of them needed support from the interviewer. Therefore, we have shown that the chosen visualization techniques are appropriate to explore such claims. We found that visualizing the number of turbines through a histogram, with the option to vary and select a variable on one of the axes, is particularly important. In addition, a map visualization is beneficial to get an idea of how the distribution occurs globally. Unfortunately, for two of the tasks **Q2** and **Q4**, only 40% and 50% of the participants could find the correct answer without help. A more dynamic grouping would have simplified the solution of these two tasks considerably, and therefore its addition would further refine the proposed technique.

**RQ3** What visual encoding can accurately capture the main visual features of a wind farm (size, spacing, terrain, nearby infrastructure) and represent them for a [quick] clear visual comparison?

Our developed visual encoding shows the relevant properties and evidences them so that most participants successfully solved the related tasks. Moreover, the design was very appealing to some participants, who found this part of the visualization the most exciting. One thing that needs to be added in the future is a legend and a scale to ensure correct interpretation.

## 9.2 Conclusion

The results of this study show that our visualization tool effectively provides valuable insights and gives the possibility to explore the EDWin dataset. Feedback from a diverse group of experts, including visualization, domain, and geography experts, helped identify areas for improvement, such as data enrichment and data quality. Additionally, suggestions were provided for further enhancing the tool's capabilities in providing a more comprehensive understanding of wind farm data, for example, by including wind conditions at the site.

If we try to answer our main research question, "What do wind farms look like?" we conclude that the multitude of wind farm designs makes it impossible to arrive at a singular conclusion. Our visualization tool demonstrates this diversity and complexity, allowing users to explore wind farms based on numerous features such as country, landform, land cover, number of turbines and turbine spacing. Those features often determine the final wind farm design. For example, the terrain conditions of a particular area can play a significant role in determining the type of wind farm design that is most suitable. Certain terrains may be more favourable for specific designs, while others may present unique challenges requiring different design solutions. The country where a wind farm is located can also significantly impact its design, as different countries may have different standards and regulations that must be adhered to. In addition, the number and type of wind turbines used and the distance between the turbines also affect the appearance of a wind farm. Therefore, the tool serves as the answer itself, providing an interactive and user-friendly way to explore the world of wind farm designs and the factors that determine them.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Figures

3.1	Wind farm layout in meters. Source: [26] . . . . .	12
3.2	Wind farms layout visualization with heat map for elevation level. Source: [27] . . . . .	13
3.3	Wind farm visualization that combines location and layout. Source: [28] . . . . .	13
3.4	U.S. Wind Turbine Database Viewer. Source: [29] . . . . .	14
3.5	Mapped: How the US generates electricity. Source: [30] . . . . .	14
4.1	Nine-stages framework . . . . .	18
4.2	Nested Model . . . . .	18
4.3	Design triangle . . . . .	19
5.1	Number of OSM nodes per country . . . . .	27
5.2	OSM data completeness . . . . .	27
5.3	OSM power tag completeness by country . . . . .	28
5.4	OSM hub height tag completeness by country . . . . .	28
5.5	OSM rotor diameter tag completeness by country . . . . .	29
5.6	OSM manufacturer tag completeness by country . . . . .	29
5.7	Erroneous records for wind turbines . . . . .	30
5.8	Turbine spacings for each turbine with neighbouring turbine . . . . .	32
5.9	Offshore wind farms with only certain sections clustered (left) and wind farms not recognised as such at all (right) when using $\epsilon = 800$ m . . . . .	34
5.10	Erroneous clustering of wind farm in China with $\epsilon = 800$ m . . . . .	34
5.11	Erroneous clustering of wind farm in India with $\epsilon = 800$ m . . . . .	35
5.12	Tool to find spatial radius for clustering . . . . .	36
5.13	Data structure and source overview . . . . .	37
5.14	A glimpse into the "shapes" of wind farms . . . . .	41
5.15	Explained variance by number of components . . . . .	42
6.1	Draft of visual encoding for individual wind farm typology . . . . .	46
6.2	Initial interface of the prototype . . . . .	47
6.3	Upper section of interface with opened filter options . . . . .	48
6.4	Interface with activated filter only showing wind farms on oceans and closed filter options . . . . .	49
6.5	Map view interface: initial view . . . . .	49

6.6	Frequency distribution view interface: Bar charts for categorical y-axis . . .	50
6.7	Frequency distribution view interface: Histogram for numerical y-axis . . .	51
6.8	Random wind farms view interface . . . . .	52
6.9	Simplified prototype architecture . . . . .	56
6.10	Detailed prototype architecture . . . . .	56
7.1	Overview of tasks and interview questions in user study evaluation . . . .	63
7.2	Random sample of wind farm configurations for <b>Q5</b> and <b>Q6</b> . . . . .	65
7.3	Task accuracy results of user responses . . . . .	65
7.4	Frequency of important Wind Farm and Turbine characteristics determined by Domain Experts . . . . .	70
8.1	Scatter plot of turbine spacing vs. power output . . . . .	78
8.2	Scatter plot of turbine spacing vs. rotor size . . . . .	78

# List of Tables

4.1	Simplified Methodology . . . . .	23
5.1	Discrete Land Cover naming convention, Source: [57] . . . . .	38
5.2	Discrete Landform naming convention, Source: [58] . . . . .	39
5.3	Final wind turbines example data structure with ratio of missing data . .	44
5.4	Final wind farms example data structure with ratio of missing data . . .	44
6.1	Color encoding for marker cluster on map view . . . . .	50
6.2	Color encoding of map elements used for individual wind farm encoding on "Random wind farms view" . . . . .	53
7.1	Frequency of Prototype feedback from users on potential improvements and additions . . . . .	71
7.2	Guidelines and components scores . . . . .	74
7.3	ICE-T Methodology questionnaire with results . . . . .	75



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.



# Bibliography

- [1] Ara Begum, R., R. Lempert, E. Ali, et al. Point of Departure and Key Concepts. In: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. In Press. p. 8-12.
- [2] United Nations. (2015). Paris Agreement. Paris: Available in [http://unfccc.int/paris\\_agreement/items/9485.php](http://unfccc.int/paris_agreement/items/9485.php).
- [3] IEA (2022), *Electricity Market Report - January 2022*, IEA, Paris. Available in <https://www.iea.org/reports/electricity-market-report-january-2022>.
- [4] IEA (2022), *World Energy Outlook 2022*, IEA, Paris <https://www.iea.org/reports/world-energy-outlook-2022>, License: CC BY 4.0 (report); CC BY NC SA 4.0 (Annex A).
- [5] Joyce Lee and Feng Zhao. *Global wind report 2021*, Mar 2021.
- [6] IEA (2021), *Renewables 2021*, IEA, Paris. Available in <https://www.iea.org/reports/renewables-2021>.
- [7] Dunnett, S., Sorichetta, A., Taylor, G. et al. Harmonised global datasets of wind and solar farm locations and power. *Sci Data* 7, 130 (2020). <https://doi.org/10.1038/s41597-020-0469-8>.
- [8] Joseph T Rand, Louisa A Kramer, Christopher P Garrity, Ben D Hoen, Jay E Diffendorfer, Hannah E Hunt, and Michael Spears. A continuously updated, geospatially rectified database of utility-scale wind turbines in the united states. 2020.
- [9] Peter Enevoldsen and Scott Victor Valentine. Do onshore and offshore wind farm development patterns differ? *Energy for Sustainable Development*, 35:41–51, 2016.
- [10] Marina Haller. *Enriched Data of Wind Farms (EDWin)*. <https://doi.org/10.5281/zenodo.7558885>, January 2023.

- [11] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.
- [12] Silvia Miksch and Wolfgang Aigner. A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers Graphics*, 38:286–290, 2014.
- [13] S.C. Bhatia. 8 - wind energy. In S.C. Bhatia, editor, *Advanced Renewable Energy Systems*, pages 184–222. Woodhead Publishing India, 2014.
- [14] Marc van Grieken and Beatrice Dower. Chapter 23 - wind turbines and landscape. In Trevor M. Letcher, editor, *Wind Energy Engineering*, pages 493–515. Academic Press, 2017.
- [15] TU CatalogPlus. TU Wien Bibliothek. <https://catalogplus.tuwien.at/>, 2013. (Accessed: 2022-11-20).
- [16] ScienceDirect. Elsevier. <https://www.sciencedirect.com/>, 1997. (Accessed: 2022-11-20).
- [17] Google Scholar. Google LLC. <https://scholar.google.com/>, 2004. (Accessed: 2022-11-20).
- [18] Sultan Al-Yahyai, Yassine Charabi, Adel Gastli, and Abdullah Al-Badi. Wind farm land suitability indexing using multi-criteria analysis. *Renewable Energy*, 44:80–87, 2012.
- [19] Yenisleidy Martínez-Martínez, Jo Dewulf, and Yannay Casas-Ledón. Gis-based site suitability analysis and ecosystem services approach for supporting renewable energy development in south-central chile. *Renewable Energy*, 182:363–376, 2022.
- [20] Felix Nitsch, Olga Turkovska, and Johannes Schmidt. Observation-based estimates of land availability for wind power: a case study for czechia. *Energy, sustainability and society*, 9(1):45–45, 2019.
- [21] Pia Nabielek. *Wind power deployment in urbanised regions : an institutional analysis of planning and implementation*. TU Wien Academic Press, Wien, 2020.
- [22] Marcus Eichhorn, Mattes Scheftelowitz, Matthias Reichmuth, Christian Lorenz, Kyriakos Louca, Alexander Schiffler, Rita Keuneke, Martin Bauschmann, Jens Ponitka, David Manske, and Daniela Thrän. Spatial distribution of wind turbines, photovoltaic field systems, bioenergy, and river hydro power plants in germany. *Data*, 4(1), 2019.
- [23] Andrew Kusiak and Zhe Song. Design of wind farm layout for maximum wind energy capture. *Renewable Energy*, 35(3):685–694, 2010.

- [24] Mahdi Abkar and Fernando Porté-Agel. A new wind-farm parameterization for large-scale atmospheric models. *Journal of renewable and sustainable energy*, 7(1):13121, 2015.
- [25] Longyan Wang, Andy C.C. Tan, and Yuantong Gu. Comparative study on optimizing the wind farm layout using different design methods and cost models. *Journal of wind engineering and industrial aerodynamics*, 146:1–10, 2015.
- [26] Souma Chowdhury, Jie Zhang, Achille Messac, and Luciano Castillo. Unrestricted wind farm layout optimization (uwflo): Investigating key factors influencing the maximum power generation. *Renewable Energy*, 38(1):16–30, 2012.
- [27] Jim Y.J. Kuo, David A. Romero, J. Christopher Beck, and Cristina H. Amon. Wind farm layout optimization on complex terrains – integrating a cfd wake model with mixed-integer programming. *Applied Energy*, 178:404–414, 2016.
- [28] Ben He, Shaoli Yang, and Knut H. Andersen. Soil parameters for offshore wind farm foundation design: A case study of zhuanghe wind farm. *Engineering Geology*, 285:106055, 2021.
- [29] The United States Wind Turbine Database (USWTDB) Viewer. <https://eerscmapp.usgs.gov/uswtodb/viewer>, 2016. (Accessed: 2022-08-09).
- [30] Mapped: How the US generates electricity. Carbonbrief. <https://www.carbonbrief.org/mapped-how-the-us-generates-electricity/>, 2017. (Accessed: 2022-08-09).
- [31] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.
- [32] Leilani Battle and Jeffrey Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Computer Graphics Forum*, 38:145–159, 06 2019.
- [33] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):491–500, 2019.
- [34] Miriah Meyer and Jason Dykes. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 08 2019.
- [35] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org> , 2017. <https://www.openstreetmap.org> (Accessed: 2022-10-25).
- [36] Overpass API. <https://overpass-api.de/> (Accessed: 2022-10-25).

- [37] OpenStreetMap contributors. OpenStreetMap contributors. About. <https://www.openstreetmap.org/about>, 2017. (Accessed: 2022-10-25).
- [38] OpenStreetMap contributors. OpenStreetMap contributors. Copyright., 2017. <https://www.openstreetmap.org/copyright> (Accessed: 2022-10-22).
- [39] OpenStreetMap contributors. Elements. OpenStreetMap Wiki. <https://wiki.openstreetmap.org/wiki/Elements> (Accessed: 2022-10-22).
- [40] OpenStreetMap contributors. Tag:generator:source=wind. OpenStreetMap Wiki. <https://wiki.openstreetmap.org/wiki/Tag:generator:source%3Dwind> (Accessed: 2022-10-22).
- [41] Python package index - requests. <https://pypi.org/project/requests/> (Accessed: 2022-10-22).
- [42] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, page 226–231. AAAI Press, 1996.
- [43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [44] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.
- [45] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, sep 1975.
- [46] Robert Gasch and Jochen Twele. *Introduction to Wind Energy*, pages 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [47] W. In Cutler J. Cleveland and Christopher Morris, editors, *Dictionary of Energy (Second Edition)*, pages 638–655. Elsevier, Boston, second edition edition, 2015.
- [48] Hoen, B.D., Diffendorfer, J.E., Rand, J.T., Kramer, L.A., Garrity, C.P., and Hunt, H.E., 2018, United States Wind Turbine Database v5.2 (October 12, 2022): U.S. Geological Survey, American Clean Power Association, and Lawrence Berkeley National Laboratory data release, <https://doi.org/10.5066/F7TX3DN0>.
- [49] Rand, J.T., Kramer, L.A., Garrity, C.P. et al. A continuously updated, geospatially rectified database of utility-scale wind turbines in the United States. *Sci Data* 7, 15 (2020). <https://doi.org/10.1038/s41597-020-0353-6>.

- [50] Joseph Dwyer and David Bidwell. Chains of trust: Energy justice, public engagement, and the first offshore wind farm in the united states. *Energy Research & Social Science*, 47:166–176, 2019.
- [51] Python package index - reverse\_geocoder. [https://pypi.org/project/reverse\\_geocoder/](https://pypi.org/project/reverse_geocoder/) (Accessed: 2022-10-17).
- [52] Python package index - pycountry. <https://pypi.org/project/pycountry/> (Accessed: 2022-10-17).
- [53] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017.
- [54] Google. Earth engine data catalog | google developers. <https://developers.google.com/earth-engine/datasets/catalog>, (Accessed: 2022-10-17).
- [55] Google. Earth engine code editor | google developers. [code.earthengine.google.com](https://code.earthengine.google.com/), (Accessed: 2022-10-17).
- [56] Marcel Buchhorn, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6), 2020.
- [57] Marcel Buchhorn, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6), 2020.
- [58] David M. Theobald, Dylan Harrison-Atlas, and Christine M. Monahan, William B. & Albano. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PLOS ONE*, 10(12):1–17, 12 2015.
- [59] Jeffrey J Danielson and Dean B Gesch. Global multi-resolution terrain elevation data 2010 (GMTED2010). Technical Report 2011-1073, 2011.
- [60] Open-elevation. <https://open-elevation.com/>, (Accessed: 2022-10-17).
- [61] Open-Elevation. Public api documentation. <https://github.com/Jorl17/open-elevation>, (Accessed: 2022-10-29).
- [62] J.P. Snyder, P.M. Voxland, and Geological Survey (U.S.). *An Album of Map Projections*. Number No. 1453 in An Album of Map Projections. U.S. Government Printing Office, 1989.
- [63] Ryan Cohn and Elizabeth Holm. Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data. *Integrating Materials and Manufacturing Innovation*, 10(2):231–244, Jun 2021.

- [64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [65] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [66] Python package index - plotly. <https://pypi.org/project/plotly/> (Accessed: 2022-10-22).
- [67] François Chollet et al. Keras. <https://keras.io>, 2015. (Accessed: 2022-10-23).
- [68] Hans-Hermann Bock. *Clustering Methods: A History of k-Means Algorithms*, pages 161–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
- [69] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748, 2020.
- [70] Guido vanRossum. Python reference manual. *Department of Computer Science [CS]*, (R 9525), 1995.
- [71] Danny Goodman. *JavaScript Bible*. John Wiley & Sons, 2001.
- [72] David Sawyer McFarland. *CSS3: the missing manual*. " O'Reilly Media, Inc.", 2012.
- [73] Anubhav Hanjura. *Heroku cloud application development : a comprehensive guide to help you build, deploy, and troubleshoot cloud applications seamlessly using Heroku*. Community Experience Distilled. Packt Publishing, Birmingham, England, 1st edition edition, 2014 - 2014.
- [74] Python package index - dash. <https://pypi.org/project/dash/> (Accessed: 2022-10-22).
- [75] Elias Dabbas. *Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python–no JavaScript required*. Packt Publishing Ltd, 2021.
- [76] Mapbox. <https://www.mapbox.com/>, 2010. (Accessed: 2022-10-23).
- [77] Bill Kastanakis. *Mapbox Cookbook*. Packt Publishing Ltd, 2016.
- [78] Python package index - dash-leaflet. <https://pypi.org/project/dash-leaflet/> (Accessed: 2022-10-22).
- [79] Numa Gremling. Turf.js – geoverarbeitung im browser. FOSSGIS e.V., 2016. <https://doi.org/10.5446/19745> (Accessed:2022-10-12).

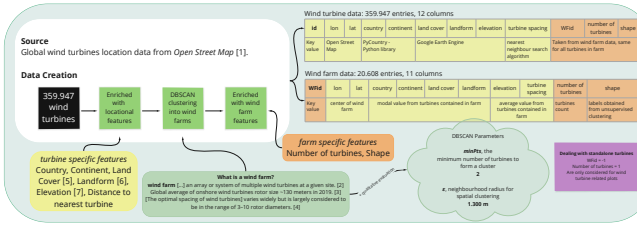
- [80] Clientside callbacks. Plotly. <https://dash.plotly.com/clientside-callbacks>, 2012. (Accessed: 2022-10-29).
- [81] Mapbox/supercluster: A very fast geospatial point clustering library for browsers and node. Mapbox. <https://github.com/mapbox/supercluster>, 2016. (Accessed: 2022-10-29).
- [82] Dash datatable. Plotly. <https://dash.plotly.com/datatable>, 2012. (Accessed: 2022-10-25).
- [83] Mapbox Studio. Mapbox. <https://studio.mapbox.com/>, 2010. (Accessed: 2022-11-11).
- [84] Chaomei Chen and YUE YU. Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies*, 53:851–866, 04 2000.
- [85] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005.
- [86] M. Tory and T. Moller. Evaluating visualizations: do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, 2005.
- [87] Michael Auer, Melanie Eckle, Sascha Fendrich, Luisa Griesbaum, Fabian Kowatsch, Sabrina Marx, Martin Raifer, Moritz Schott, Rafael Troilo, and Alexander Zipf. Towards using the potential of openstreetmap history for disaster activation monitoring. In Kees Boersma and Brian M. Tomaszewski, editors, *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. ISCRAM Association, 2018.
- [88] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, Mar 2016.

## Appendix Prototype Manual



# What do wind farms look like? Manual for wind farm visualization prototype

## Data



**Terminology**  
wind farm size - number of turbines contained in farm  
feature - wind farm characteristic e.g. land cover, elevation

## Example wind farms



## Example Tasks

Which land form is predominant for the installation of windfarms in Austria?

How do the landforms upper slope and summit compare in terms of turbine spacing?

Where do you find the biggest wind farms and how do they look like?

## Functionalities

This is an approved original version of this thesis available in the TU Wien Bibliothek. The original version of this thesis is available in the TU Wien Bibliothek.

### Filter

Visualizing Wind Farms 30,677

Country: [Dropdown], Continent: [Dropdown], Land Cover: [Dropdown], Landform: [Dropdown], Shape: [Dropdown], Elevation: [Dropdown], Distance to nearest turbine: [Dropdown]

Map: [Map view]

### View 2: Frequency Distribution

Obtain the frequency distribution of wind farms and turbines over a selected feature.

Clicking on legend items will filter and requery data.

### View 1: Map Visualization

Get an overview on how wind farms and turbines are distributed in the world.

Zoom in to see details of a specific wind farm.

### View 3: Random Wind Farms

Get a feeling on how wind farms that match the filtered features look like.

Click on a random wind farm to get information about single wind farms.

## Cheatsheet feature values

Country	Continent	Land Cover	Landform	Shape	Distance to nearest turbine	Elevation	Number of turbines
138 countries the world	Africa Antarctica Asia Europe North America Oceania South America unknown	Agriculture Closed forest Herbaceous vegetation Herbaceous wetland Moss and lichen Oceans and seas Open forest Permanent water bodies Shrubs Snow and ice Sparse vegetation Urban unknown	Flat Lower slope Summit Upper slope Valley Waterbody unknown	Single turbine Polygon Not clustered Less than 5 turbines Lines Irregular lines	Min: 10 meters Max: 13.150 meters	Min: -46 meters Max: 4.684 meters	Min: 1 (standole) Max: 3.296

**Sources**

[1] Open street map. <https://openstreetmap.org/>. [Online] Accessed: 2022-10-02.

[2] Ottler J., Cleveland, Christopher Morris, Dictionary of Energy (Second Edition), Elsevier, 2015, Pages 638-655, ISBN 9780080968117 <https://doi.org/10.1016/B978-0-08-096811-7.50023-8>.

[3] Mochumitha Jaganmohan (2020) Average rotor diameter of onshore wind turbines worldwide from 2000 to 2025 [Infographic]. Statista. <http://www.statista.com/statistics/1085649/onshore-wind-turbines-average-rotor-diameter-globally/> [Online] Accessed: 2022-10-02.

[4] D. Nett, S., Sorichetta, A., Taylor, G. et al. Harmonised global datasets of wind and solar farm locations and power. *Sci Data* 7, 130 (2020). <https://doi.org/10.1038/s41597-020-0469-8>

[5] Bonnafant, M.; Lesiv, M.; Tsendbazar, N. - E.; Herold, M.; Bertels, L.; Smets, B. Copernicus Global Land Cover Layers-Collection 2. *Remote Sensing* 2020, 12 Volume 108, 1044. [doi.org/10.3390/rs12061044](https://doi.org/10.3390/rs12061044)

[6] Trumbald, S. D. M., Harrison-Atlas, D., Monahan, W. B., & Albano, C. M. (2015). Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PLoS one*, 10(2), e0143619

[7] Multi-resolution Terrain Elevation Data 2010 courtesy of the U.S. Geological Survey

