Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

CICLO XXXIV

Settore Concorsuale: 01/B1 - INFORMATICA
Settore Scientifico Disciplinare: INF/01 - INFORMATICA

# Applied Deep Learning and Data Science with a Human-Centric and Data-Centric Approach

*Presentata da:*
**Luca Casini**

*Coordinatore Dottorato*
**Prof. Daniele Bonacorsi**

*Supervisore*
**Prof. Marco Roccetti**

ESAME FINALE ANNO 2023

This dissertation is submitted for the degree of Doctor of Philosophy
in Data Science and Computation

February 2023

This rippling uncertainty
beneath our bones is still the
true state of all things

---

*P. Elverum*

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. This dissertation contains fewer than 50,000 words including appendices, bibliography, footnotes, tables, and equations and 60 original figures.

February 2023

# Acknowledgements

Doing a PhD is without any doubt the hardest thing I ever did in my life. Getting through it would have been impossible without the support of many people who I would like to thank now.

First I want to express my gratitude to my supervisor Prof. Marco Roccetti for setting me on the path to become a scientist and pushing me forward at every step. A special mention then goes to all the scholars with whom I had the honor to collaborate in my research activity. In particular Givoanni Delnevo, Valentina Orrù and Nicolò Marchetti.

I want to thank Prof. Bob Sturm, Joris, Laura, Nicolas and Marco for making my stay in Stockholm, taking part in the MUSAiC project, as pleasant as it was. Working with you has been an endless source of inspiration and I look forward to meeting you again.

Finally I dedicate this thesis to all my friend, old and new, that supported my along this journey, often unknowingly. To my parents for giving me the chance to follow every interest I have had, making it possible to even be here. And to Serena for being the best partner I could hope for.

# Abstract

The term Artificial intelligence acquired a lot of baggage since its introduction and in its current incarnation is synonymous with Deep Learning. The sudden availability of data and computing resources has opened the gates to myriads of applications. Not all are created equal though, and problems might arise especially for fields not closely related to the tasks that pertain tech companies that spearheaded DL.

The perspective of practitioners seems to be changing, however. Human-Centric AI emerged in the last few years as a new way of thinking DL and AI applications from the ground up, with a special attention at their relationship with humans. The goal is designing a system that can gracefully integrate in already established workflows, as in many real-world scenarios AI may not be good enough to completely replace its humans. Often this replacement may even be unneeded or undesirable.

Another important perspective comes from, Andrew Ng, a DL pioneer, who recently started shifting the focus of development from "better models" towards better, and smaller, data. He defined his approach Data-Centric AI.

Without downplaying the importance of pushing the state of the art in DL, we must recognize that if the goal is creating a tool for humans to use, more raw performance may not align with more utility for the final user. A Human-Centric approach is compatible with a Data-Centric one, and we find that the two overlap nicely when human expertise is used as the driving force behind data quality.

This thesis documents a series of case-studies where these approaches were employed, to different extents, to guide the design and implementation of intelligent systems. We found human expertise proved crucial in improving datasets and models. The last chapter includes a slight deviation, with studies on the pandemic, still preserving the human and data centric perspective.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The research activity conducted in the four years that led to this thesis has been characterized by a horizontal approach to the discipline of data science, which resulted in the application of the same concepts to a variety of different fields. There is however a common thread running through all these projects, which is the idea of putting the role of human experts and users at the center of the design process.

This philosophy is often referred to as *Human-Centric (or Human-Centered) AI*, as opposed to work that too narrowly focuses on the models and algorithms while loosing track of the relationship the finished model will have with its users and their goals. The insight humans provide if involved can, in turn, help in improving the models, evaluating the results, and refining the data, which is the prime driving force behind data science applications and machine learning systems. The attention to data, which should be "good" rather than just "big", has shaped the *Data-Centric AI* movement which, similarly to the human-centric approach, is opposing the focus on scaling model size. At the intersection of the two we found the value of using domain expertise to clean and reduce datasets, as well as evaluating the model outputs and help in debugging. This Human-AI collaboration loop results in more performant, more reliable and overall more usable systems.

In the next section will go over the research background and introduce the questions we will address in the following chapters. Finally, we conclude this introduction by providing an outline of the dissertation.

## 1.1 Research Background

In 2018 the Association for Computing Machinery (ACM) nominated Geoffrey Hinton, Yann LeCun and Yoshua Bengio as the three recipients of the prestigious Turing Award for "conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing" [1]. In a way, this was an official recognition of the dawn of a new era of Artificial Intelligence (AI), spearheaded by adoption of deep learning (DL).

It has been 66 years since the infamous Dartmouth Conference [2] that characterized the golden age of AI, whose popularity rose and waned multiple times over the years, and while many fundamental questions are yet to be answered we can clearly see how today AI seems to be much more of reality than it was before. The advancements of deep learning, which some define as a fourth industrial revolution, have been made possible by the increase in computational power and the enormous quantity of data made available by the internet and the ubiquitous presence of computing devices like smartphones.

In this scenario, Data Science and machine learning have become such a fundamental part of computer systems that people capable of wielding these powerful tools become highly sought after [3]. As Barreto hinted at back in the 1990s, neural networks could be considered a new programming paradigm. One where the data replaces the explicit instructions given in code by the programmers [4].

It is interesting to notice, however, how in the world of traditional software development we have seen a number of sub-fields emerging as computers become more and more integrated in our society. For example, tools needed to be created to help programmers organize their work, keep track of changes and enable collaboration with others; various testing methodologies became a necessity when systems needed to be dependable and run constantly; similarly the discipline of human-computer interaction emerged as computers entered homes and offices to become tools at the disposal of non-technical audiences [5].

When it comes to this new programming paradigm instead, the corresponding disciplines that address the same concepts are still taking shape and lack the same depth. This could be only a matter of time, and tools to address the engineering task of developing machine learning and deep learning systems are being introduced more and more. But when people deal with AI there is also a fundamental difference with classical software development, somewhat of a philosophical stance: the idea that, ultimately, an intelligent system should be completely capable of doing tasks like we humans do without the need of intervention on our part. Many believe that a general artificial intelligence (AGI) is right around the corner [6] but the current success of deep learning, with models capable of surpassing human

performance on many difficult tasks [7, 8] and even breaking new ground on problems like protein folding [9] is still resulted in narrow AI that can only do specific things. It is true that size and performance seem to scale linearly [10], but very big and flexible models, recently dubbed foundation models, like OpenAI's *GPT-3* [11] or Deepmind's *Gato* [12] still have quite visible limits and it is reasonable to argue that maybe more data and computation is not enough, also considering how training those systems is inaccessible to most.

Whether or not AGI is near, many scholars are warning about the danger of not considering these technologies in relation to humanity and advocate for the development of systems that are human compatible, to cite the title the last book by Stuart Russell [13]. Rather than focus on artificial intelligence we should strive for what Minsky at al. defined humanistic intelligence: "[...] intelligence that arises because of a human being in the feedback loop of a computational process, where the human and computer are inextricably intertwined" [14].

When we shift the focus from raw performance to actual utility for the users and we start treating deep learning as yet another tool in the data science box a lot of interesting research considerations start emerging from the limitations of these new technologies. In particular, we are interested in the human-centric (or human-centered) AI movement, exemplified by groups like the HCAI team at Stanford , and, relatedly, to the data-centric perspective introduced by Andrew Ng [15].

Human Centric AI, as the name suggests, means that humans should be put at the center of AI research rather than algorithms and benchmarks [16]. This shift in perspective invests all aspects of AI models, from the inputs they receive for training, to the way their outputs are presented. It is easy to draw a parallel to human computer interaction from the software engineering world but in this case interacting with models could mean a number of different things: interacting with the data during the preparation of the dataset and design of the model, interacting with the development of the model, or interacting with the trained model and its predictions.

With Data Centric AI instead the focus is on data as the most effective source of improvement now that the algorithmic and coding part of deep learning has been "solved" as Ng says in an interview introducing the concept [15]. He highlights how scaling up model size may still be worthwhile for some applications but in some fields that amount of data is simply not available. In these situations, moving away from "big" data towards "good" data is the most effective strategy. Quantifying what good data means is obviously very dependent on the context and on the task at hand, but it is easy to see how human experts can play an active and invaluable role in providing this measure.

If we think about the life-cycle of an AI model, from its design to its productization and use, we can identify a number of phases where human input could be integrated, injecting their values and perspective. In the literature, this design choice is called human-in-the-loop AI [17, 18]. In this context humans can be included in a downstream way, the most common, or in an upstream way.

The first approach corresponds to what is sometimes called active learning or interactive machine learning [19] and focuses on the most obvious loop in AI-human interaction, that is inputting new data and getting a prediction. Here a model makes a prediction that a human in turn uses to make a decision while also providing feedback to the model, often by providing a ground truth that may have been wrong or noisy in the training set.

The latter instead arises by considering the development loop of the model. Here ML engineers and domain experts can work together and iteratively improve both on the model architecture and the dataset [20]. The human role here can be that of identifying the best dataset and weeding out uninformative or misleading examples or that of providing an evaluation of results which are difficult to judge in a purely mathematical way (e.g., art, text, moral decisions) . This kind of loop is crucial in all the situations in which there is ambiguity in the data, perhaps when the dataset was not explicitly build for the purpose of training a model, or ambiguity in the question the model is trying to answer, where bias may be lying in plain sight but needs human expertise to be spotted.

When companies, especially smaller ones, are trying to implement practical AI systems the issue of data quality is as big of a problem as data quantity [21]. This obstacle to actionable AI can realistically only be solved through human intervention when the data its already there

Having humans and AI work together is also about the way things are presented. This concerns performance, outputs but also the internal structure of the model itself. Careful choice of metrics and data visualization can help in this regard.

Metrics can obfuscate a lot of the shortcomings of a model if they are not chosen carefully and even more importantly, they should be aligned with the values of the users. Failing to recognize this is a form of the famous McNamara Fallacy and can lead to catastrophic results when the model is released in the wild. A high scoring model for classification might give the impression that it has learned a meaningful representation of the data but looking at what the model considers reveals that this is not the case. For example, Ribeiro et al. found that distinguishing huskies from other dogs was done by looking at the background and Sturm et al. found that music genre recognition models were not considering

musical features that humans relate to said genres [22, 23]. This problem is what in turn enables adversarial attacks to deep learning models and prevents their use in certain high stakes situations [24].

Visualization techniques can enable debugging a model when things go wrong or just simply allow us to understand what goes into making a prediction [25, 26]. Furthermore, they can increase the value associated with a model prediction by making it more interpretable and digestible to non-technical audiences.

On the topic of interpretability, we also need to consider situations in which humans are the recipients of a decision coming from an AI system. In situations where there is a lot at stake and failure cannot be tolerated even for human beings, people will tend to not trust machines and will require an explanation of how it arrived at its conclusion [27–29]. In this context, the performance of the model is the least of the concerns while legal and ethical questions take the main stage.

While there is an increasing corpus of research on explainable AI, researchers like Cynthia Rudin advise against the use of post-hoc explanation of black box models in supporting law enforcement and the justice system in the USA [30]. She noticed how dataset in those situations were very biased, and arguably the society from which the data came from is even more, and claims that black-box systems hid away those biases behind a curtain of good predictive performance.

In the world of healthcare, we see that automated diagnoses are not well received by patients, who prefer to interact with a human physician [31]. At the same time, while intelligent systems can achieve superhuman results in visual diagnostic exams like CT scans, the combined performance of humans and AI is even higher [32, 33]. Furthermore, most of the time caring for a patient and forming a diagnosis goes beyond the correct interpretation of a single exam and entails high-level abstract reasoning that machines do not excel at.

## 1.2 Research Questions

Given the contextual framework described in the previous section, let us now go over the research questions we will address in this thesis through the discussion of the cases studies reported in the following four chapters. As anticipated, this thesis contains applications of data science to very different fields, but all of these applications are informed by the same approach to the discipline.

The following research questions serve to highlight these themes across the different chapters:

**RQ1: How can human expertise be integrated in the development of an AI system?** As we discussed in the background section, human expertise can be integrated at different levels of the machine learning pipeline. In each of the projects that compose this thesis we will highlight how we leveraged human expertise and how much it contributed to the performance of the final model.

**RQ2: Can an under-performing system still be useful?** "All models are wrong, but some models are useful" states an infamous maxim by George Box that every data scientist knows. However, there are cases in which a model is performing worse than what one would expect in a similar application. The reason for this can be many but, as we said, performance metrics are not the only indication of value. We will investigate how a model can be helpful for humans even without breaking records in benchmarks.

**RQ3: Are metrics always sufficient to evaluate a model?** Metrics give us an indication of the performance of a model and, together with the loss function, encode the goal of a deep learning model, or, in other words, what question it is answering. Sometimes, however, the question is not easily encoded by a simple mathematical function, and we need to involve domain experts in the evaluation.

**RQ4: What is the role of visualization in human-centric applications?** Visualization techniques are a powerful tool to discover relationship in the data but can also play an important role in the development of a deep learning model by providing a view of what the model is learning and doing. These tools are even more valuable when coupled with domain expertise as non-trivial concepts can emerge from the right plot.

**RQ5: When are classical approaches preferable to deep learning?** Deep Learning has replaced other techniques as the state of the art in virtually any field of application. However, there are situations in which classical methods from statistical learning can be more useful. In the context of this thesis, we found this to be the case for studies concerning the pandemic, when the lack of data meant training neural networks and making predictions was difficult but there was still insight to be gathered from other types of models.

## 1.3 Thesis Outline

We conclude this introductory chapter by including a structured summary for the rest of the thesis, to provide an overview of every chapter at a glance:

**Chapter 1** serves as the introduction, describing the aims and contents of the work and giving a brief account of the research background that help framing all methods and conclusions.

**Chapter 2** describes a project which involved building a system for predicting the failure of mechanical water meters for an Italian energy company. Along the way many considerations related to our research question emerged, especially of data quality and of Human-AI collaboration.

**Chapter 3** deals with the design of a model able to automate the remote sensing phase of an archaeological expedition. This is the phase in which archaeologists look for potential sites of interest using satellite and aerial imagery. This is a perfect example of Human-AI collaboration as automating this extremely time-consuming task would only help archaeologist focus on more important parts of their work while still requiring their involvement in validating the outputs. We used two different approaches, both leveraging transfer learning and showed how refining the dataset and data representation together with domain experts can improve performance.

**Chapter 4** illustrates the creation of a model for automatic generation of symbolic music in the style of traditional Irish and Swedish folk. Differently from the other field of applications we depicted, music is characterized by the absence of a clear-cut notion of a "good" output. This makes the roles of human experts in the loop fundamental, as it allowed us to overcome design difficulties and create a model able to beat the previous state of the art. In the final part we also discuss how such a model can be integrated into a music co-creation workflow.

**Chapter 5** contains a number of observational studies performed during the pandemic, highlighting the themes of Research Question 5. In a situation where understanding what is happening is as important as obtaining good predictions, classical models can be a better choice compared to a black box like neural networks.

**Chapter 6** finally concludes the thesis, summing up the common insights gather across all the different research projects.

# Chapter 2

# Automatic Detection of Defective Water Meters

This chapter describes the results of a collaboration project with a company that provides various services in norther Italy like waste management, electricity, natural gas, and water supply [34–41]. The water supply division of the company wanted to use artificial intelligence methods to solve a problem with water meters they had been facing for a while. For obvious privacy reasons the company will remain unnamed and the relevant information about the dataset and models will be discussed without revealing details connected to their business.

This project posed some interesting challenges that highlighted the importance of involving domain experts in the design of machine learning systems as they hold the key to the good interpretation of data. After an initial phase of bad performance, we inquired the managers and engineers in the company and with the insight they provided we were able to refine the dataset and obtain a model with good performance.

In a further experiment with new data, we found that the inclusion of additional information in the form of categorical features regarding the nature of each water meter and its use degraded performance instead of improving. We hypothesized the cause for this was connected to the curse of dimensionality and inspired by the distribution of the categorical attributes (a Pareto-like distribution), we tried different techniques to reduce dimensionality like binning and principal component analysis, albeit without solving the issue. Thus, we devised an unusual approach, in which we filter the dataset accordingly and create a model for each categorical attribute which can be used as an ensemble with better overall accuracy.

In the final part of this collaboration, it emerged the theme of how performance metrics in classification problems can be misaligned with the actual costs and benefits of a company. While we were never informed on the actual details and numbers, we proposed the company to adjust the classifier threshold in a way that made sense for their business process, even if it were not the mathematical optimum.

## 2.1 Detecting Faulty Devices

In this section we will describe the process that led to creation of the first classifier for faulty devices, including discussing background and related works and detailing the neural network we designed for the task.

### 2.1.1 Background

Let us start by discussing some related work. Anomaly detection is a common task in the world of machine learning, and it is characterized by a small number of interesting data points that are different from the rest in some way. Usually, the approach consists in learning the distribution of the ordinary data and comparing the anomalous instances to that. If there is enough data on the axis of time, one could also compare each individual instance to its own past to see if there is any deviation from the norm.

In the case of our specific task, however, we have a very heterogeneous dataset when it comes to its properties (e.g., consumption, use, meter type, etc.) and, more importantly, with huge differences and irregularities in the frequency of the readings. This made classical approaches to anomaly detection more complex to adapt to our problem and pushed towards an end-to-end solution using neural networks.

Additionally, while we are plenty of papers in the literature for individuating anomalies like a leakage or a failure, in water distribution pipelines, there is a not surprising scarcity of papers that discuss methods for detecting anomalies in water meters [42–44].

Those that do seem to focus on the use of simple heuristics and statistical analyses. For example, Roberts and Monk developed a simple algorithm that individuates possible anomalies, occurring at a given water meter, when a decreasing trend in water consumption is observed along a series of readings which is updated just quarterly [45]. Monedero et al, instead, propose an approach to detect tampering activities in mechanical water meters that employs a very basic statistical analysis for identifying [46]: either a low rate in water consumption, a sudden

stoppage of that consumption, or simply a decreasing consumption trend.

For sure, the advent of electrical water meters, along with telemetry that can provide water consumption readings on a per hour basis, could significantly alter this picture and allow the use of more common online anomaly detection techniques that are already employed in managing other types of utilities like electricity [47].

### 2.1.2 Dataset Description

The company had no specific database put in place for this task, so a significant part of this project consisted in working together to define a proper dataset to train our machine learning algorithms on. Our starting point is a huge database of reading and metering devices which contains a lot of redundant information and is used mainly for administrative and billing operations. The data spans a period of 4 years, from 2014 to 2018, with around 15 million rows, representing individual readings that pertain to over 1 million metering devices, including both working and defective.

Table 2.1 list the 14 attributes associated with the individual readings in the starting database. Attribute 1 is used as the primary key for grouping the readings for each device. Table 2.2 lists the 17 attributes associated with each device, with attribute 17 representing our target variable. As we will discuss later, not every attribute was deemed useful for our purpose of training a predictive model.

Table 2.1: Attributes in the readings database

| No | Attribute name | No | Attribute name |
|---|---|---|---|
| 1 | Water Meter ID | 8 | Reader ID |
| 2 | Reading ID | 9 | Type of Contract |
| 3 | Reading Value | 10 | Reading Validity |
| 4 | Reading Date | 11 | Certification on the ERP |
| 5 | Prev. Reading Value | 12 | Final Billing |
| 6 | Prev. Reading Date | 13 | Reason for Reading |
| 7 | Reading Frequency | 14 | Accessibility |

### 2.1.3 Dataset Cleaning and Pre-processing

The preprocessing phase is a critical part of any data science project. In this specific instance we had to invest significant work into cleaning up the format of each column and then extracting the relevant information from the starting database by matching readings to their respective device and making sure everything was

Table 2.2: Attributes in the devices database

| No | Attribute name | No | Attribute name |
|---|---|---|---|
| 1 | Water Meter ID | 10 | Installation Date |
| 2 | Producer Code | 11 | Plant |
| 3 | Producer Description | 12 | Type of Contract |
| 4 | Material ID | 13 | Geographical Zone |
| 5 | Material Description | 14 | Accessibility |
| 6 | Max/min Reading Value | 15 | Use Category |
| 7 | Meter Type ID | 16 | Address |
| 8 | Meter Type Description | 17 | Operation (Faulty/Non Faulty) |
| 9 | Year of Construction | | |

coherent. We devised, and revised, a set of rules together with the domain experts which we referred to as a "semantic of validity" which we will now describe. After that we proceeded with the well-established practices of standardizing the numeric variables and of transforming categorical data into one-hot encoded vectors.

### Feature Selection

Selecting features is a key task, since irrelevant or redundant features can impact the training activities [48, 49]. Numerical Features were easy to choose as they only consist in the water consumption values (difference between current reading and the previous) to which we combine the number of days since the last reading to help with their irregular frequency (Attributes 3, 4, 5, 6 of the readings dataset).

Further, on the basis of precise suggestions provided by the company, we also included the following additional features from the devices dataset: producer (Attribute 2), material (Attribute4), meter type (Attribute 7), and use category (Attribute 15).

Table 2.3 reports all the aforementioned selected features.

Table 2.3: Final set of features used to train the model

| # | Features | # | Features |
|---|---|---|---|
| 1 | Reading Value | 5 | Serial Number of the Producer |
| 2 | Previous Reading Value | 6 | Material ID |
| 3 | Reading Date | 7 | Meter Type ID |
| 4 | Previous Reading Date | 8 | Use Category |

### Semantics of Validity

Many of the readings that compose the datasets came with numerous inconsistencies and impurities, whose causes depend on organizational conflicts between different business processes. The company obviously never gave us much detail on those processes but allowed us to interact with domain experts who helped us define what we call a "semantic of validity".

The first rule of this semantic was considering attribute 10 (Reading Validity). This is set by a human operator once he reads a value on the meter and validates its correctness. Experts suggested that readings labeled as explicitly non-valid should not be taken into consideration. Table 2.4 reports the number of non-valid measurements with respect to the total.

Table 2.4: Readings: Valid/Non-valid (attribute 10)

| Attribute 10 | of Readings |
|---|---|
| Initial | 15,129,379 |
| Non-valid | 1,898,128 |
| Valid | 13,231,251 |

In addition to attribute 10, we were suggested to consider also attributes 11 (Certification on the ERP) and 12 (Final Billing), as their combined values offer a stronger indication of correctness. Their meaning is as follows:

(i) has been (correctly) read/collected on site by a human operator,

(ii) has been (correctly) recorded onto the company ERP system,

(iii) has been (correctly) billed to the final client.

While attributes 10, 11, and 12 can take a total of 45 different combinations, just 7 of those cover almost 99% of the total amount of readings in the dataset. These specific combinations are shown in Table 2.5. Some of those, even if still "valid", refer to different administrative aspects of the readings, however, to create our dataset we are only interested in the readings that reflect the true numbers shown on the device. The experts pointed us to the first combination in the Table, with Attribute 10 equal to 1 and Attribute 11 and 12 equal to 2. For the sake of simplicity, we will refer to those readings as those enjoying the *1-2-2 Factor*.

Up until this point we only talked about readings as we were dealing with the rows of a database containing all of them. But since we need to predict meters, we had to aggregate the readings by device. Table 2.6 shows how many meters, that also have the *1-2-2* Factor, possess at least a certain number of readings, from 1 to

Table 2.5: Readings: main categories (with relative amount of readings)

| Attribute 10 | Attribute 11 | Attribute 12 | Readings |
|---|---|---|---|
| 1 | 3 | 2 | 407,592 |
| 1 | 2 | 4 | 282,527 |
| 1 | 2 | 6 | 132,409 |
| 1 | 2 | 5 | 110,363 |
| 1 | 2 | 3 | 106,742 |
| 1 | 3 | 5 | 105,957 |
| | Other | | 229,079 |
| | **Total** | | **13,231,251** |

5. Additionally, the second-to-last row contains the total number of meters, while the last one is the total amount of faulty meters.

Unfortunately, of the total amount of readings we now considered valid, some were mathematical adjustments of the readings, estimated values of presumed water consumption values computed for billing purposes. We asked the domain experts to point them out as their presence interfered with phenomenon we wanted to detect. Table 2.7 shows the quantity of readings we had to discard.

At this point we thought that, given our "semantic", the dataset was clean enough to paint a sensible picture of the phenomenon and so we tried training a model. When tested on a holdout set with newer data however, we obtained an AUC score of 0.61. Consequently, we furthered the discussion with company experts to find what the cause might be.

Our intuition is that that semantics disregards the role played by time. In

Table 2.6: Number of meters with the 1-2-2 Factor

| 1-2-2 Factor | Meters |
|---|---|
| 1-2-2 ($\geq 1$) | 1,154,054 |
| 1-2-2 ($\geq 2$) | 1,091,334 |
| 1-2-2 ($\geq 3$) | 1,038,337 |
| 1-2-2 ($\geq 4$) | 981,420 |
| 1-2-2 ($\geq 5$) | 915,441 |
| Faulty ($\geq 1$) | 23,752 |
| **Total** | **1,239,977** |

Table 2.7: Proportion of real vs. adjusted measurements (with the 1-2-2 Factor)

| Type | # of Readings |
|---|---|
| Real | 8,185,163 (69%) |
| Adjustments | 3,671,419 (31%) |
| **Total** | **11,856,582** |

particular we found out that the time at which each reading is taken can vary wildly between different devices and even between successive readings of the same one. Figure 2.1 exemplifies this phenomenon. The $x$-axis represents the difference in consumption, in terms of cubic meters of water, between two subsequent readings enjoying the 1-2-2 Factor, while on the $y$-axis we can see the time intervals (measured in days) between two subsequent readings with the 1-2-2 Factor.

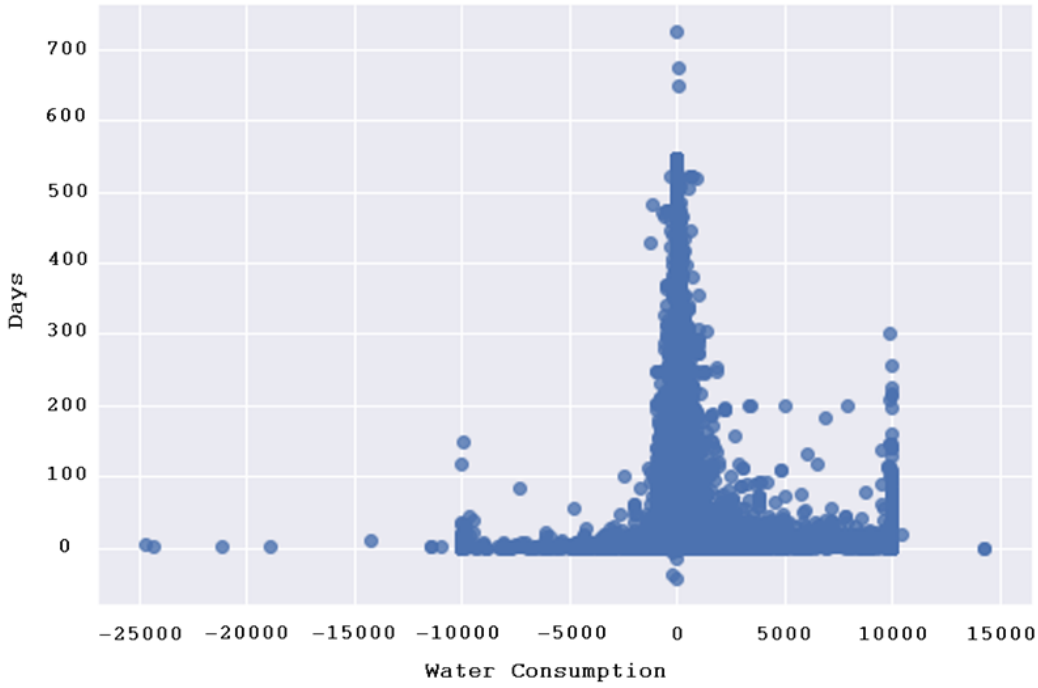



Figure 2.1: Time intervals vs differential water consumption (two consecutive readings)

Looking at the $y$-axis we can see that there are a number of readings that are extremely separated in time. This is so extreme that even if we considered them valid it would take years to have more than one reading. In any case those

are suspect as the Italian law prescribes at least two or three readings per year, depending on the type of use. We also see some outliers when it comes to the difference in consumption. Those may depend on a series of factors according to the company, such as the necessity to reset them to 0 (hence the points that lie on $y = 0$ as the operation is registered immediately), which we were not interested in, however. All these considerations pushed us to reconsider the semantics we previously defined and to introduce additional constraints to guarantee coherent readings when it comes to consumption and time.

A reading is to be considered valid only if all the following requirements are satisfied:

(i) a human operator has read a certain reading value at the reading site.

(ii) that reading value has been correctly recorded onto the company ERP and billed to the client

(iii) time and consumption difference with the previous reading, following rule 1 and 2, are what we can considered coherent.

For the sake of simplicity, we will call this enhanced semantics as the *X-Factor*. Notice that the enforcement of the X-Factor to our initial dataset makes the number of valid readings fall down to less than two million. On this reduced but clean dataset we trained the model described in the next Section.

### 2.1.4 Model Description

We designed a deep neural network specifically for our task. Figure 2.2 contains a schematic representation of it. The architecture is characterized by the presence of two parallel branches that are responsible for processing and extracting features from the time-series data for consumption and the categorical attributes for each meter. The model was implemented using the *Keras* library that is built on top of Google's *Tensorflow* framework for deep learning [50, 51].

The time-series branch uses a recurrent neural network, specifically the Gated Recurrent Unit (GRU) variant, followed by a fully connected layer with ReLU activation [52, 53]. The choice of a RNN allows the model to process an arbitrary number of readings without any problem (apart from the necessity of padding) but we soon realized with the company that requiring more than 5 readings makes the model difficult to put in production. The categorical branch instead features a simple stack of fully connected layers, also with ReLU activation, that takes in the one-hot encoded vectors. Both branches feature a dropout mechanism to avoid overfitting [54]. The output tensors of the two branches are then concatenated into a single one that goes through a final dense layer of the same size before the

final binary output that computes the classification probability with the softmax function. The model is optimized using stochastic gradient descent (SGD) applied to the binary cross-entropy loss.

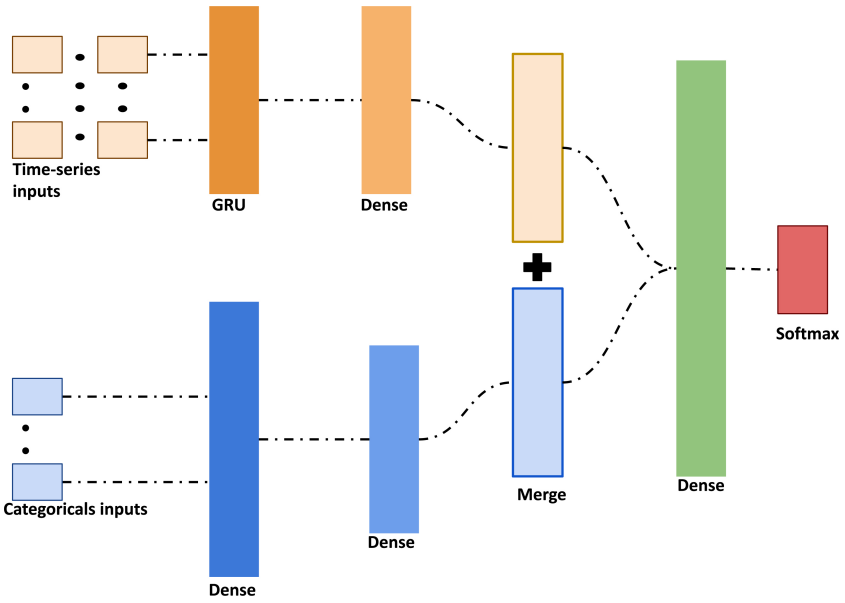



Figure 2.2: Schematic representation of the artificial neural network we designed. The top branch (in yellow) processes the time series for water consumption. The bottom branch (in blue) processes the categorical attributes. The resulting features are concatenated before the last layer.

As already seen in table 2.6 the dataset is highly imbalance as the fault we are trying to detect is, fortunately, quite rare. This is a common problem in this kind of scenario. We addressed it by oversampling the minority class using synthetic data created with the algorithm SMOTE-NC [55, 56]. This is a variant of classic SMOTE that can deal with categorical data as well. We used the area under the Receiver Operating Characteristic curve (AUC-ROC) as the performance metric to report results and compare different models during development [57].

### 2.1.5 Results

We trained the model with meters whose readings, following the semantic of validity, were taken in the period from the beginning of 2014 to mid 2018. The dataset comprised around 45,000 non-faulty meters and around 15,000 faulty meters. As anticipated, to help in the learning process, we used SMOTE-NC to over-sample the faulty water meters in the training achieving a balance between the classes. One tenth of the data was used for validation and one fifth for testing.

Training and validation performance for our neural network is reported in Table 2.8, showing results using at least a certain number of readings for each meter (from 2 to 5). With an AUC score ranging from 0.82 to 0.88 in the validation phase we can confidently consider our experiments successful. The company deemed the models working with devices with at least 2 or 3 readings the most appropriate. These struck a balance between accuracy and usability, given the time necessary to accumulate 4 readings or more can be too long. Following this, we performed an additional and final testing experiment on a new testset with around $30,000$ devices with new readings with the X-Factor gathered in from the middle to the end of 2018. The test yielded an AUC score of 0.86 and 0.89, using 2 or 3 readings respectively, confirming the efficacy of our model and semantic.

Table 2.8: Validation and Testing performance for the neural network model with different number of readings as inputs.

| Input Readings | AUC (validation) | AUC (test) | Additional Test |
|---|---|---|---|
| 2 | 0.90 | 0.82 | 0.86 |
| 3 | 0.93 | 0.85 | 0.89 |
| 4 | 0.95 | 0.87 | |
| 5 | 0.97 | 0.88 | |

Additionally, we have conducted a comparative analysis with some common machine learning algorithms. This is important as artificial neural network are known to be under-performing on tabular data, which part of our dataset is, even though this is mainly true for smaller dataset [58, 59]. We experimented with all the following traditional learning algorithms, implemented in the open-source library `scikit-learn` [60]:

- Linear Regression (LR),
- Lasso (LA),
- Elastic Net (EN),
- Classification and Regression Tree (CART),
- K-nearest neighbors (KNN),
- Adaptive Boosting (AB),
- Gradient Boosting (GB),
- Random Forest (RF),
- Multi-Layer Perceptron (MLP, with one hidden layer with 100 neurons).

The bar plot in Table 2.9 shows the AUC scores for each of these models, to be compared with our deep neural network (DNN) with 3 readings in the rightmost bar. Some models, like the GB and MLP ones, get close to a 0.80 score, which we could consider acceptable. Nonetheless, in this specific application deep learning seems to be the way to go, with results that are consistently higher.

Table 2.9: AUC scores for different machine learning models against our neural network using 3 readings

| Model | LR | LASSO | EN | KNN | CART | AB | GB | RF | MLP | DNN |
|---|---|---|---|---|---|---|---|---|---|---|
| AUC (test) | 0.79 | 0.78 | 0.65 | 0.61 | 0.66 | 0.67 | 0.80 | 0.79 | 0.79 | 0.85 |

## 2.2 Making Categorical Data Helpful

Although the company deemed the results shown in the previous section satisfactory, we wanted to extend the collaboration and further validate our methods with new additional data. Thus, they provided us new dataset pertaining a different group of water meters with at least 3 readings that follow the semantic we defined. It contained 17,714 devices, of which 15,652 were non-defective and 2,062 were defective.

The company informed us that the process behind the labeling of faulty devices was slightly different this time, without going into much detail. However, the change was transparent to us as we still got a binary flag for good and bad samples. Still the performance might have been affected by this change.

We found that our model best model, retrained on this new dataset was underperforming, with an AUC score of 0.78. Trying to understand what could have gone wrong, we noticed that removing categorical data altogether from the inputs actually resulted in performance closer to the original experiments. Consequently, we hypothesized that the problem lies with the so-called curse of dimensionality, as categorical attributes are one-hot encoded and result in a total of 205 additional dimensions.

In this section we describe how we tried to leverage this additional information in order to further increase performance, first trying dimensionality reduction techniques and noticing the Pareto-like distribution of data and finally using a non-traditional approach consisting in using the categorical attributes as filters for different models.

### 2.2.1 Dimensionality Reduction

Our dataset contains both categorical and numerical features. Most machine learning algorithms are estimated through the optimization of a real-valued function in a continuous space and thus can directly work on numerical data as they are (or at most some preprocessing is required for numerical stability).

Categorical variables on the other hand do not clearly translate to real-valued space where a function can be optimized and usually require some special treatment. The most common approach is called one-hot encoding and consist in creating a $n$-dimensional vector where $n$ is the number of categories for each feature.

It is easy to see how the size of this vector can grow very fast, making the solution space very big and sparse, and possibly damaging the performance of a model [61]. The problem of exploiting categorical variables while avoiding or mitigating this downside is common to any machine learning approach and is

actively studied with many different techniques used to address it [62–64].

All of them follow similar ideas of generating new features that encapsulate the information of the original high-dimensional representation but in a more compact space. These new features should be as few and uncorrelated as possible, while providing most of the information contained in the original data. Hopefully the loss of information given by the procedure is balanced by the increased performance of the resulting model that can be more easily fit to the dataset.

Commonly used for numerical data, Principal Component Analysis (PCA) is the most famous technique for reducing dimensionality [65]. It re-projects data points in a new, smaller, space determined by the largest eigenvectors that account for most of the variance in the data, making it mathematically similar to singular value decomposition (SVD). The same idea applied to categorical data is called Multiple Correspondence Analysis (or MCA). Essentially, it performs the same re-projection using SVD but working on a contingency table [66–68].

A generalization of MCA is CATegorical Principal Components Analysis (or CATPCA) which however is equivalent in the case of one-hot encoded variables like our own [69].

Dimensionality can be also reduced through Multi-Dimensional Scaling (MDS) technique. MDS aims to represent observations in lower-dimensional metric space while trying to reproduce the original distance between them in best way it can using a non-linear transformation [70, 71].

Using a latent variable model, some researchers try to find groups and clusters in latent space and use those as features for classification [72–74].

In our case however we would expect our neural model to do this implicitly and train a new latent variable model would incur in the same problems. A simple, yet crude, way to reduce the number of dimensions in certain situations it the technique called Binning. As the name suggest we can group categorical attributes in bins that share common characteristics, with the classic example being age brackets [75]. The other common scenario is to create a category for uncommon values. Obviously, this method is quite a lossy process and the decision to group certain values together should be informed by domain expertise or some mathematical analysis. However, if the information loss is relatively small it can be a good solution. A similar operation is often done for numerical data and is called Censoring.

As we will see in the next section, classical methods did not seem to provide any advantage in our case. Thus, instead of using categorical features as input to the model we employed them as a driver for data selection, thus eliminating, from the start, the need for a dimensionality reduction of the categorical space. The

resulting models have a narrower scope but combined together they yield a higher accuracy.

### 2.2.2 Pareto Distribution of Data

To investigate the difference in performance between the model with and without the categorical data we started by plotting the distribution of values for each categorical attribute.

In Figure 2.3, we can see four histograms showing the distribution of values for each categorical variable in our model, which we will call A, B, C, and D. These can take, respectively 98, 45, 48 and 14 different values. For each categorical variable, we have a dotted curve with the cumulative percentage distribution of those n values over our devices. For the sake of conciseness, the figure only shows the distribution for defective devices as the respective histograms for non-defective meters would show very similar results. It is quite evident that the shape of the distribution for all the categorical variables in the dataset is that of a power law, commonly known as a Pareto distribution [76].

The popular principle connected to this type of distribution is that 80% of the population exhibits only 20% of the possible values, while the remaining 20% features the other 80%. In other words, a few values are highly representative of the whole group while the majority is quite rare. In the general framework of power-law distribution the actual values can be obviously different from 80 and 20. In our case we see that around 90% of the dataset is covered by around 20% of the attributes.

Table 2.10 summarizes this finding reporting the total number of values and the amount necessary to cover the majority of the dataset for each categorical variable.

| Variable | N. of Values | Most Frequent | N. of Meter Devices | |
|---|---|---|---|---|
| | | | Defective | Non-defective |
| A | 98 | 23 (23%) | 1855 (90%) | 13474 (86%) |
| B | 45 | 7 (16%) | 1854 (90%) | 13707 (88%) |
| C | 48 | 11 (23%) | 1889 (92%) | 13369 (85%) |
| D | 14 | 3 (21%) | 1945 (94%) | 13963 (89%) |

Table 2.10: A quasi-Pareto distribution of the categorical characteristics

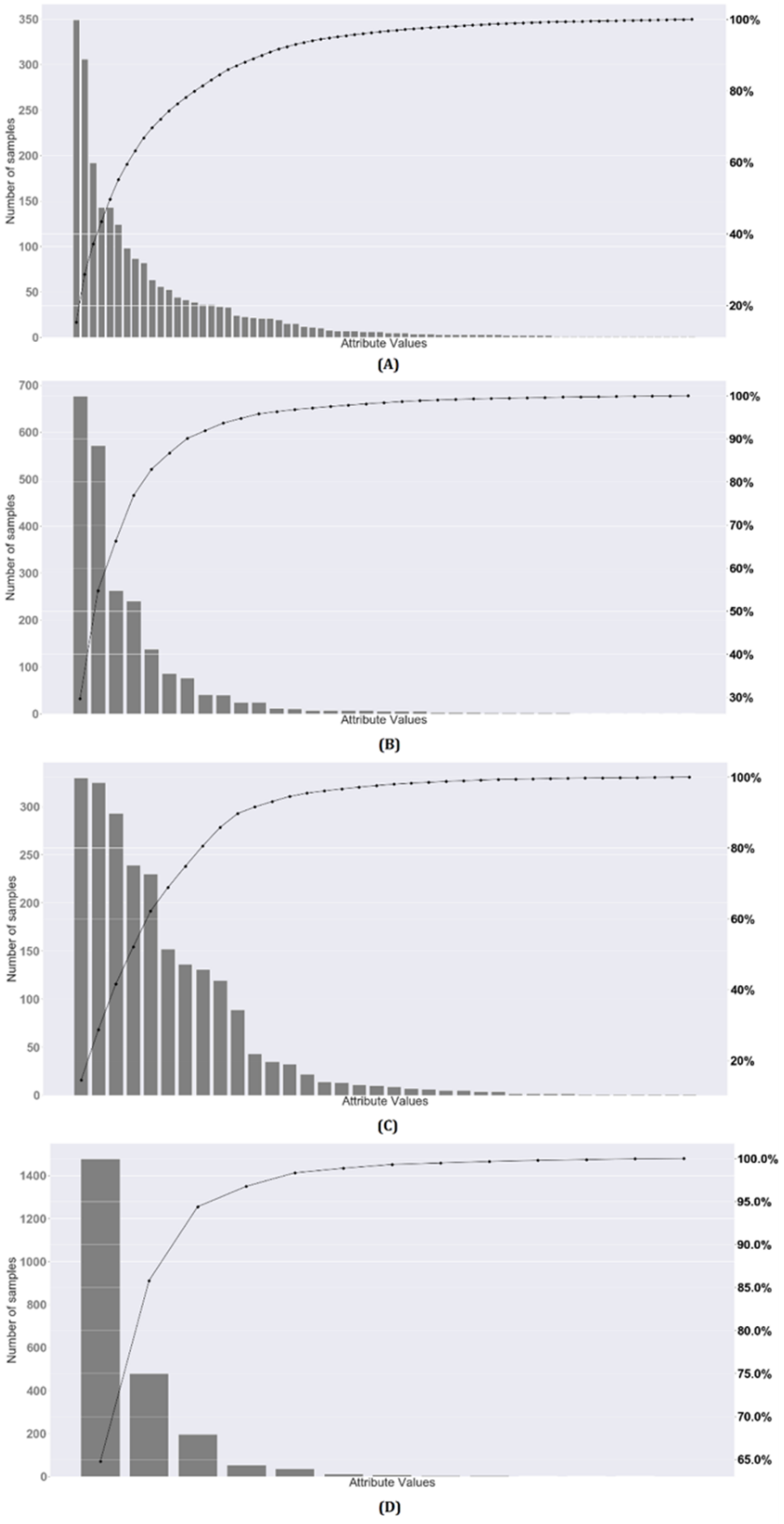



Figure 2.3: Histograms of value counts for each categorical variable

### 2.2.3 Results

In this section we report the results from the initial tests we conducted and the subsequent experiments where we tried reducing the dimensionality with the techniques described before. Table 2.11 lists the size of the categorical space and the performance of each model. The dataset used for all models discussed here used 2062 defective meters and 15,652 non-defective meter.

Table 2.11: results for the four experiments with categorical data

| Exp | Dimension | AUC-ROC validation | AUC-ROC testing |
|---|---|---|---|
| #1 | 0 | 85% | 86% |
| #2 | 205 | 78% | 83% |
| #3 | 128 | 73% | 81% |
| #4 | 48 | 76% | 85% |

In the first experiment the input included only the numerical values of the readings without any categorical attribute. Instead, in the second experiment we used both the readings and the categorical values mentioned above (A, B, C and D) transformed with one-hot encoding.

In the third experiment we tried reducing the dimensionality with PCA [65]. Using it on one-hot encoded data makes it equivalent to MCA. We selected the first 128 principal components, which accounted for approximately 90% of the variance in the dataset.

Lastly, in the fourth experiment we tried employing binning. The size of the bins is informed by the analysis described in the previous section. The resulting categorical space given by one-hot encoding has 48 dimensions in total: 44 given by the sum of the most frequent categorical values plus 4 for the "other" bin added for each categorical feature.

Unfortunately, none of the techniques we tried was effective in increasing the performance of the base model with no categorical attribute. Binning is slightly improving the situation in testing but not in training while PCA seems to be damaging in every situation. Since the domain experts in the company believed that these attributes should contain some useful information that we could exploit we decided to try a completely different approach before giving up on their use.

### 2.2.4 Using Categorical Data as a Filter

Since using categorical data as additional information to provide to the model did not seem to be helpful in this situation, we wanted to try another approach to

the problem, considering also how the resulting model could be included in the company processes.

Our idea was that of using the categorical inputs as filters for the dataset rather than inputs. We can create datasets that contain only devices featuring the most frequent values for a certain categorical device and train different models for each of them. These models would be less general but hopefully more accurate in that specific cases. When the inputs do not fall in the "filtered" categories we considered then we can just fall back on the general model.

Table 2.12 shows the results of these models trained on the filtered dataset according to attribute A, B, C or D, which we called respectively $DLM_A$, $DLM_B$, $DLM_C$, and $DLM_D$.

Table 2.12: Validation and test result for the four attribute-specific models

| Model | AUC validation | AUC test |
|---|---|---|
| $DLM_A$ | 0.87 | 0.88 |
| $DLM_B$ | 0.87 | 0.87 |
| $DLM_C$ | 0.86 | 0.88 |
| $DLM_D$ | 0.87 | 0.87 |

The resulting models from this procedure can also be used together, in an example of what are called ensemble methods and in particular of the bagging technique [77]. The predictions from the models can combined by majority voting or simply averaged. This approach can be used for example whenever we have a device that possesses the characteristics of all the four models together. Obviously, this can work only for a limited quantity of meter devices, yet it could provide finer predictions, whenever applicable.

Table 2.13 shows the results the performance obtained by combining the predictions on the test set from the four models and using the voting strategy. This applies to the intersection of the four groups which counted 2304 non-defective devices and 313 defective ones. As shown, we have an improvement in AUC-ROC that increased to the value of 90%.

Table 2.13: Testing results using an ensemble made from all the categorical models

| Model | PPV | NPV | TPR | TNR | F1(P) | F1(N) | Acc. | AUC |
|---|---|---|---|---|---|---|---|---|
| DLM/Bagging | 65% | 96% | 61% | 83% | 63% | 95% | 91% | 90% |

### 2.2.5 Comparison with Other Methods

Like in the previous section, we wanted to compare the model with other machine learning algorithms that could perhaps work better on the tabular data. We used the implementation available in the python library scikit-learn, using the default parameters for each of them. The methods used are SVM, CART, GB and RF. The last two especially seemed to be the best performing in the previous experiment and are known to be good for this type of data. Table 2.14 shows the results.

Table 2.14: Results for different machine learning algorithms

| Model | AUC validation | AUC testing |
|---|---|---|
| SVM | 0.69 | 0.80 |
| CART | 0.65 | 0.73 |
| GB | 0.72 | 0.89 |
| RF | 0.78 | 0.89 |

All methods seem to have difficulties learning in this context, but we can see that Random Forest has almost the same performance as our neural network. Surprisingly, they score even higher in the testing dataset which could be an artifact of the dataset itself. Since it was sampled from the company, we unfortunately have no way of knowing if the process was biased. When we performed a cross-validation experiment, however, AUC scores showed a high variance for RF that is not present for the neural network, thus we can conclude that our model is more robust.

## 2.3 Integrating AI in a Human Decision Process

In this section we discuss how our model can be integrated in the existing human processes that are already established in the company. Those consideration arose at the end of our collaboration, when we suggested a few modalities of use of such models. While the details may only apply to this specific instance of applied machine learning the general idea can be carried over to other contexts in which a decision must be made and that present imbalance in the dataset and an asymmetry in the importance of certain mistakes over others.

### 2.3.1 Decision Thresholds in the X-Factor Model

In the previous section we have shown how, with the help of human experts in the company, we created a usable dataset starting from a huge database of readings and attributes created for accounting purposes and then used it to train a deep learning model that was able to detect the type of faulty devices the company was looking for. Now we will discuss how this classifier relates to the established processes in the company and how it integrates into those depending on the decision threshold.

Let us start by describing how the company dealt with the problem before (to the best of our knowledge as not many details were provided for obvious reasons). The first step is the compilation of a list of candidate faulty devices by looking at their consumption history. This is done with a heuristic that considers the presence of a series of consecutive readings indicating null consumption, usually two or three. At that point, a manual process starts where analysts will manually sift through the candidates to select the devices that will be checked in person by technicians and eventually replaced if needed. Null or decreasing consumption could be due to other things obviously. Sometimes is possible to combine multiple data sources together, like natural gas consumption if the client receives both, to compare the values and exclude false positives. However, not every device is replaced due to multiple business and practical reasons (e.g., meter is inaccessible, replacement is too costly at the moment, etc.). To give some perspective on the matter here are some statistics we received based on the estimates the company makes: Every year, on average, around $10,000$ devices are considered faulty candidates; of those, almost $5,500$ are shortlisted after checks mentioned above, while around $1,500$ are finally replaced.

Now as an example let us consider the model and test dataset described in section 2.1.2, containing around $30,000$ devices with readings collected in the period mid 2018–end 2018. Of those $30,000$, $6,652$ devices were suspected as faulty, while $22,634$ were labeled as non-faulty, as stated by the company. Remind our classifier was able to make predictions in that context with an AUC score of 0.86.

Without considering anything but performance metrics, we could set the decision threshold to 0.46, the one that minimizes the number of both false negatives and the false positives, selected using the intersection point shown in Figure 2.4 and make errors as those exemplified by the confusion matrix of Figure 2.5b.

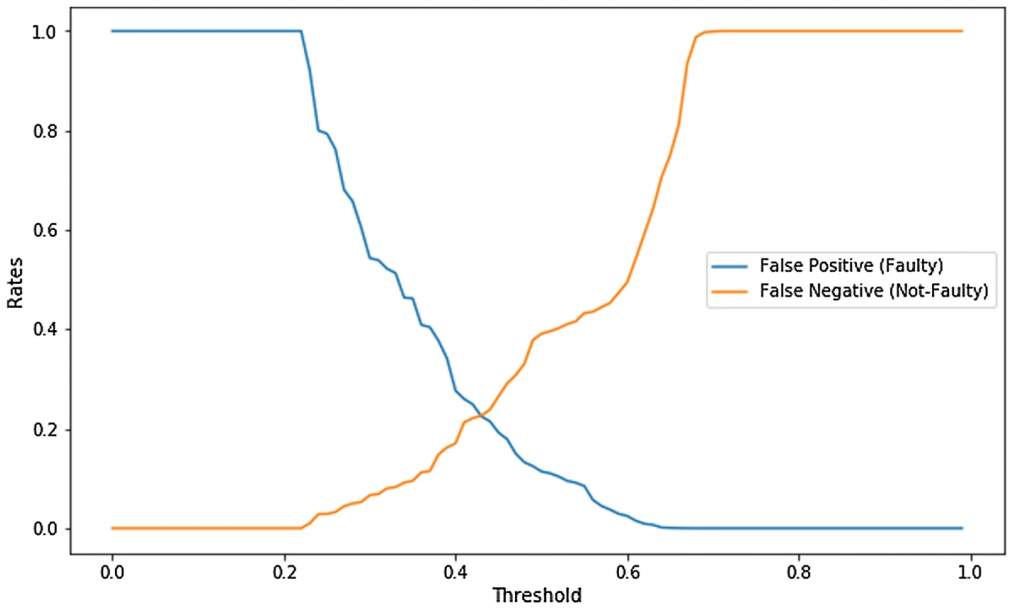



Figure 2.4: False positive rate and false negative rate depending on the choice of decision threshold

Using this model, the company could decide to ignore the devices predicted as not faulty (20, 513) and either change all the ones predicted as faulty, knowing that half of those could be actually working correctly or instead check on them manually using procedures similar to those already in place, with the added benefit of having save the time necessary for creating the shortlist. The first road is only viable if replacement is not costly. While we do not know for certain, we can imagine is not desirable for the company.

The second road instead would make them work on 8, 773 meters, for just six months, when they usually deal with only 5, 500 shortlisted devices every year, making it hardly justifiable. Additionally, with "optimal" threshold we would have 1, 937 faulty meters that are not detected. We do not know the cost of false negatives for the company but moving the decision threshold towards the direction of minimizing their number, for example with a value of 0.3 as in Figure 2.5a, would have the effect of decreasing the number of faulty meters that are never detected down to 443 while also yielding an overwhelming 18, 509 water meters to be replaced or verified.

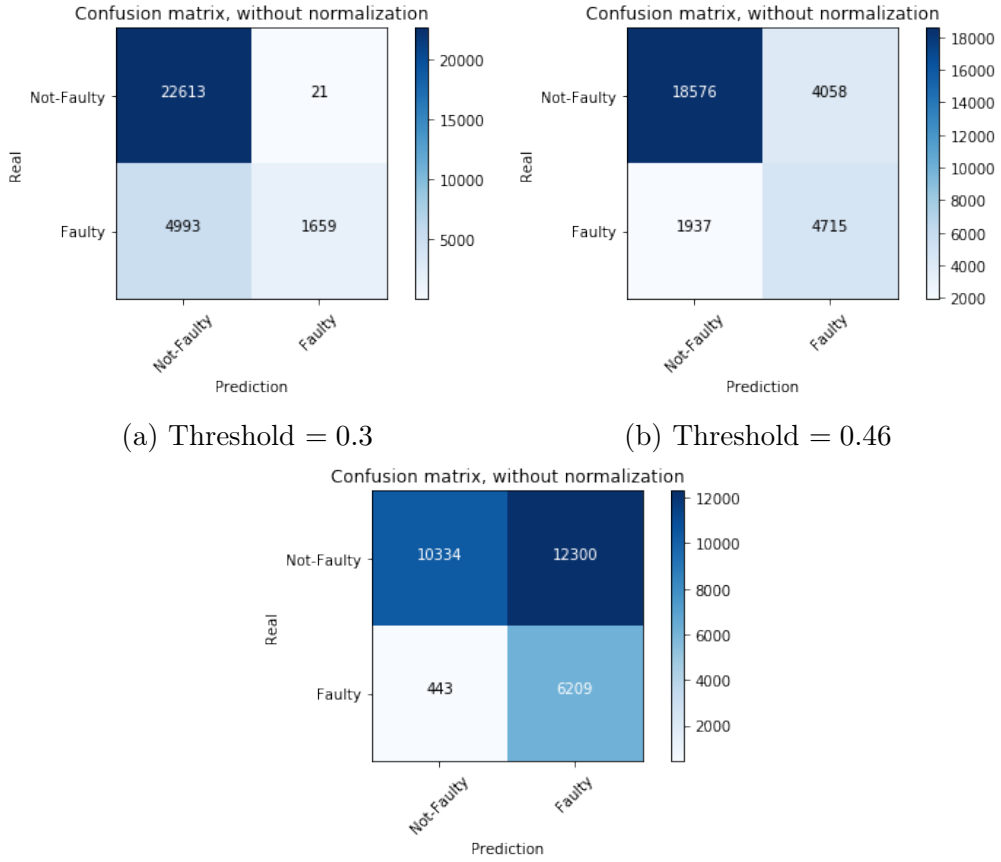
The mathematical optimum in this case is not aligned with the most useful action for the company as the value of a classification error is not neutral [78]. Integrating their perspective into the process we can instead set a higher threshold using a value like 0.65 as shown in Figure 2.5c. This approach would result in a number of suggested faulty meter of $1,680$, which is way closer to the actual number of meters that are replaced by the company with their heuristic manual method. However, we can do this without losing time on heuristics, shortlisting and any other further check. There is small cost in replacing relatively few (21) working meters which presumably do not constitute a problem. On the other hand, this strategy is potentially missing on a certain number, perhaps considerable, of faulty devices. To mitigate this, it would also be possible to find an additional threshold, slightly lower than 0.65, and manually check on those devices which are above this new value but below the former. This threshold could be chosen so that the expected fraction of previously false negative devices to check manually is around the numbers the company can already manage.

### 2.3.2 An Improved Strategy with the Categorical-filtered Model

When considering the new development described in Section 2.2 and the consequent models trained on each categorical attributes the ideas from the previous sections still stand. On the practical level things would only get a bit more convoluted.

1. We should first consider if a device possesses the categorical characteristics of either the variable A, B, C or D.
2. If that is the case, we can make a prediction using the corresponding model (either $DLM_A$, $DLM_B$, $DLM_C$, or $DLM_D$).
3. If that is not the case, we can fall back on the generic model with the X-factor and only consider the time series.

However, it should be noticed that the likelihood that a device does not possess any of those characteristics, at least in the context of the dataset we have studied, is quite low, i.e., below 10% on average, as our Pareto analysis has demonstrated. Anyhow, the use of multiple models at the same time, with some strategy to combine their predictions has the potential to make the model even more performing as section 2.2.4 showed.




(a) Threshold = 0.3

(b) Threshold = 0.46

(c) Threshold = 0.65

Figure 2.5: confusion matrices for different threshold

### 2.3.3 AUC Scores and the "Best" Model

There is one element we did not consider in the previous discussions of thresholds and performance which is linked to importance, or cost, of certain decisions or classification. In particular, the cost of each mistake and the utility of each correct answer may not be equal. In the case of binary classification, like the one we are facing, there are only two types of mistakes: false positives and false negatives. The risk connected to those two may very well be unequal, but the company did not disclose any details about it at the time of our collaboration. However, we can give a theoretical overview of how to integrate this additional constraint in the use case we described.

The technique comes from the union of decision theory from machine learning and expected utility theory from the field of economics [79, 80]. We can start by associating and score to each of the four quadrants in the confusion matrix that

correspond to the utility or cost of that type of outcome. We can assume that correctly identifying a faulty meter has a high utility if the fault makes the company lose a lot of money; conversely, the cost of sending technicians and eventually replacing a non-faulty device could be higher and undesirable. The nice thing about this approach is that those scores can be totally arbitrary and also adjusted independently of the classifiers being used.

Once we defined our utility scores, we can combine them with the prior probability of each classification outcome. This quantity can be simply estimated by considering the number of faulty devices over the total.

$$r_p = \frac{N}{N+P}; \quad r_n = 1 - r_p \tag{2.1}$$

We end up with the following formula for expected utility:

$$E[U] = u_{TP} \cdot TPR \cdot r_p + u_{FN} \cdot (1 - TPR) \cdot r_p + u_{FP} \cdot FPR \cdot r_n + u_{TN} \cdot (1 - FPR) \cdot r_n \tag{2.2}$$

Finally, using the equations below we can obtain a set of lines in the AUC-ROC space that correspond to the same level of expected utility, with slope $m$ and intercept $b$. At the intersection between the ROC curve and the line corresponding to the highest expected utility we find the best classifier according to our needs even if it may be one with a lower overall AUC score.

$$TPR = m + u_{FN} \cdot FPR + b \tag{2.3}$$

$$m = \frac{u_{TN} - u_{FP}}{u_T P - u_F N} \cdot \frac{r_n}{r_p} \tag{2.4}$$

$$b = \frac{E[U] - u_{FN} \cdot r_p - u_{TN} \cdot r_n}{u_{TP} - u_{FN} \cdot r_p} \tag{2.5}$$

Notice how only the intercept is related to the value of expected utility while the slope entirely depends on the prior probabilities and the utilities scores. In a visual way we can imagine the common ROC graph with the curve corresponding to a classifier. Using these formulas, we will have a series of lines with the same slope $m$ corresponding to different levels of the same expected utility that intersect our curve at specific points. At those points we can find the classification thresholds that correspond to that expected utility. The intersection point with the line with highest intercept represents the threshold for that maximizes utility. Moreover, if we have more than on classifier, we can plot the ROC curve for both and select the one that gives us the best utility level, even if the absolute value of the AUC curve is less or equal.

## 2.4 Conclusion

We described the results of a year-long collaboration with an Italian utility provider that operates in the norther part of the country. The project focused on the creation of a predictive model that would be able to identify water metering devices with a particular type of fault before it resulted in a complete stoppage of the appliance.

Section 2.1 documents the first experiments we performed and describes the important process behind extracting a proper dataset from the company billing database and the definition of a semantic of validity to enable us to clean up all the impurities it contained. This process would have not been possible without a collaboration with the domain experts working the company at various levels. Administrative knowledge was necessary in order to understand which attributes and rows to keep from the database. Technical expertise on the other hand was important in selecting the appropriate features that could be connected to the type of fault we were trying to model. This is a perfect example of the human-in-the-loop design process this thesis wants to portray and the increase in performance perfectly answers **Research Question #1**: The integration of human expertise is the crucial element in improving the performance of our model, guiding the choice of models and, more importantly, shaping the dataset so that it contains only the information that is most relevant to the learning algorithm.

In Section 2.2 we saw how in an additional experiment at the end of the collaboration project we had to deal with the very strange case of categorical data negatively affecting the performance of the model. Answering **Research Question #4**, this time data visualization was the key to understanding the nature of categorical data, which appeared to be following a Pareto-like distribution. Together with our intuition of data scientist pointed us in the direction of dimensionality reduction techniques to address the problem. Unfortunately, the classical approaches would not yield any significant advantage but trying the unusual approach of using categorical data as a sort of filter to split the dataset and train multiple proved successful and opened the possibility for a new way of interacting with the model through the use of multiple combined models.

The theme of how to interact with this classification model are discussed in Section 2.3, where we showed how there are various way the company could integrate our model in their decision process. Our collaboration terminated before we could see how this turned out so we can only speculate on what strategy would be in their best interest. Nonetheless, the conclusion we draw are quite general and can be applied to any situation in which there is a classification model to be inserted in a decision process. In particular we have seen how, the choice of a

decision threshold that is not the mathematical optimum may be a better choice for the company as it more easily aligns with their already established practices. This relates to **Research Question #2**. Moreover, we can notice how if the model is developed as a support to humans, rather than a substitute, a strategy could also be that of using the model classify the more obvious (as in the outputs with higher probability) data points while leaving a certain amount to be manually checked. On similar note, relating to **Research Question #3**, we also show how the metric we used, the area under the ROC curve, can be misleading in situations where the outcomes of classification are not equiprobable, and mistakes have different real-world costs associated to them. We propose a way of addressing this by using a technique coming from economics called expected utility which can be combined with then ROC curve to select the best model beyond the single AUC metric.

# Chapter 3

# Remote Sensing for Archaeology

In this chapter we will explore a completely different field of application, archaeology, tackling a task known as remote sensing. In simple terms, it consists in finding potential sites from satellite images. While very far from the application in the previous chapter, this project still evokes the same concepts of human-in-the-loop design and data-centric considerations on the quality and meaning of our training samples.

In particular, we had to deal with a dataset that was, once again, not built purposely for machine learning and with the added problem of a small number of labeled examples (few hundreds to a few thousands). Furthermore, this labeling tended to be quite imprecise, in a way that is not problematic for the manual work of humans but that can compromise the training of a model. Involving archaeologists was thus important from the start to understand how to treat the few examples we had, and, more importantly, what to consider as a negative example in a supervised learning setting.

Moreover, once a model was trained we needed their help once again as the evaluation procedure in this context is more complex than in other settings. In fact, the area the model should be used on is the same area where the examples come from and there is a considerable chance that some potential sites were indeed missed by previous surveys, especially given the size. In the event the model predicts something as a new archaeological site, then we have to decide whether it is a mistake (as an automatic testing procedure would) or if it is indeed a new site.

In the next section we will start by going over some background information and relevant research. Section 3.2 will describe the first attempt to solve this problem with a small dataset coming from the *QADIS* project and using a tile

classification approach. Section 3.3 will instead show a second attempt based on semantic segmentation models using a bigger dataset resulted from the FloodPlains project. Finally 3.4 concludes the chapter.

## 3.1 Background

We collaborated with colleagues from the department of archaeology that work in huge areas in the Near East [81–83]. Their interest in is a particular type of archaeological sites called *Tell* which is prominent in floodplains like Mesopotamia. The Arabic word Tell literally means "hill" and indicates a stratification of buildings mostly made of mud-bricks and debris that, over time, resulted in an actual artificial hill. Given the nature of the Mesopotamian floodplain, these elements tend to emerge visibly from the landscape and to be recognizable from satellite imagery. The shape, size and color can vary considerably, but they generally present an elliptic form with red and brown hues (as they are composed of clay).

These archaeologists' workflow involves a phase of "remote sensing" in which part of the large investigation area is surveyed through satellite or aerial imagery, often coming from multiple sources, and combined with old maps and reports, with the goal of spotting the contours of candidate sites and pinpointing their location.

Remote sensing can refer to any techniques and technology that involves the use of data gathered by sensors or cameras mounted on satellites or aircraft or various nature. These data can then be used to learn something about, or monitor the state of, some points of interest. This can range from keeping track of forests' growth or consumption, to the movement and size of glaciers, to even identifying pools in a neighborhood, or looking for archaeological sites, like in our case [84–89].

The emphasis on data beyond common optical imagery is due to the wide range of sensors that can be employed to highlight different properties of the target, like temperature or chemical composition, for example. This hyper-spectral remote sensing can be used in archaeology to help highlight sites that may appear more clearly in other parts of the electromagnetic spectrum [90, 91]. There is also the possibility of working with three-dimensional data collected by LIDAR technology. This requires collecting point cloud data through drones that fly at low altitude. The end results can be extremely precise, but are obviously not as easy to obtain as simple 2-dimensional photos collected by satellites [92, 93]. The analysis of all these sources of data was commonly carried out by a human expert, who knew what to look for and where, as computer vision solutions were not always viable. Doing all of this manually is obviously time-consuming but is also very important to the preparation of a mission in a faraway country that can last weeks. The recent development of deep learning for computer vision however is changing the scenario

and we now see automatic classification, detection and segmentation models based on neural networks [94]. In this context, our collaboration wanted to investigate the possibility of automating this process and to study with what accuracy a system would be able to perform the same task when provided with similar satellite imagery.

Contrary to what one could expect the archaeologists explained how the most valuable thing for them was not achieving high accuracy but rather the considerable time savings. This is because the points of interest that they find through remote sensing are checked and discussed by humans anyway before deciding whether to actually visiting them or not. A system like this could speed up the discovery phase and let them concentrate on the decision making. Additionally, the number of sites in these regions is so high that the fact of missing some of them is not considered an issue. Since the prediction would be reviewed by the domain experts in any case, overtime the system could be retrained with an ever-refined dataset and become even better. All this makes our task a perfect application of human-in-the-loop machine learning.

In next sections we will describe how we used the same satellite images used by our fellow archaeologists (other data sources like maps and old aerial photos were too low quality compared to our inputs to be combined) and approached the task using classification and segmentation techniques as seen in the literature. Also, given the size of the dataset and the limit of our computational resources, transfer learning was an important part of these project.

## 3.2 Tile Classification Approach for the QADIS Project

This section details the first approach we tried in our collaboration with the archaeologists. At the time our colleagues were working on the QADIS project, surveying an area of $1,830\ km^2$ in the *Qadisiyah* region in Iraq [95, 96]. The next section details the dataset they provided and the system we designed for the task. After that we discuss how we envisioned a possible interaction between the domain experts and the model.

### 3.2.1 Dataset

For these experiments, the archaeologists provided a small dataset of shapes corresponding to 145 confirmed sites they surveyed. Additionally, we had the shapes for 21 areas found not to be actual sites, for a total of 166. They also provided a set of 415 points coming from a previous survey by Adams [97].We decided to not consider those however, as they lacked the shape information contained in the others making it difficult to assess their size without visual inspection. Figure 3.1 shows the investigation areas and all the sites.

Starting from the satellite photos, coming from ESRI, and the shapes we were provided with we first defined a rectangle to inscribe the QADIS area. This rectangle was then divided into tiles, corresponding to images of $299 \times 299$ pixels, preserving their native resolution and roughly corresponding to an area of almost $180 \times 180$ meters. Examples of this tiles are shown in Figure 3.2. Not to lose any information, we also used a kind of intermediate tile, between two consecutive tiles, essentially obtained shifting the window by half a tile, when we extracted the images. All this resulted into a set of approximately $300,000$ tiles. The choice of size is dictated by the use of a pretrained model we will discuss later.

To be considered is the fact that while the number of tiles in this dataset is huge, it is also extremely skewed in favor of tiles representing non-sites: in fact, the number of tiles that could represent true sites is just $3,280$. To be noticed, again, is the fact that these $3,280$ refer to tiles that can also have almost a null intersection with a true site (that is, they overlap just for a very small portion with a true site); hence making the validity of the information contained in that tile arguably significant. To alleviate this problem, we selected tiles as representative of true sites, just in these two cases:

(i) they overlapped with true sites for a geographical extension of at least 10% of their area,

(ii) they overlapped with true sites for an extent of at least 30%.

In the first case, we got some 2,211 tiles, while in the second case we got some 1597 tiles out of the total amount of 3,280 tiles. These tiles represented the positive examples on the basis of which we constructed our datasets for training our deep learning model. To the positive cases, we added some negative cases (non-sites) to get two well-balanced datasets, respectively constructed of 4,422 and 3,194 examples. Those were randomly selected from tiles that had no intersection with the sites as suggested by the archaeologists.

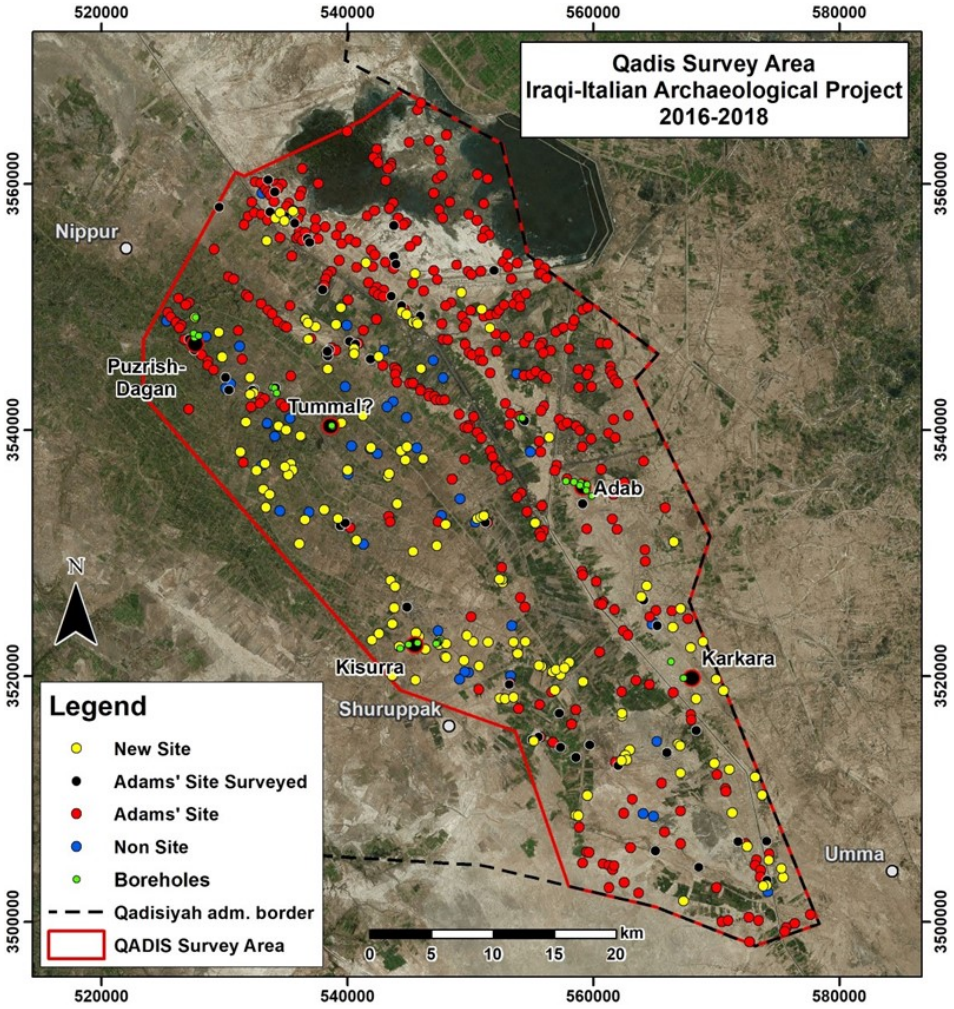



Figure 3.1: The investigation area for the QADIS project. Yellow dots correspond to the sites we used in the training set. Blue dots are the 21 sites that were mistakenly identified as such. Red dots are the 415 sites by Adams that we did not use as they did not come with a shape.

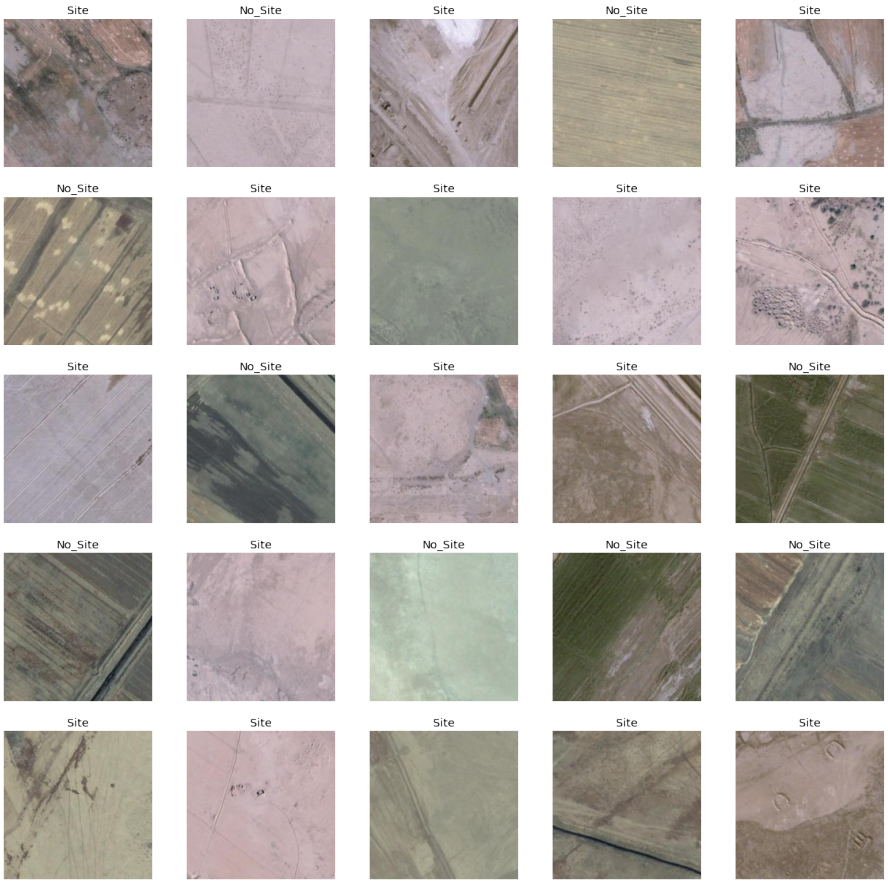



Figure 3.2: Example tiles from the dataset.

### 3.2.2 Model

Given the very small size of our dataset we considered the use of *Transfer Learning* as a viable strategy to adopt in our context. The basic idea is to start from a deep learning model that has already learnt how to classify images, and then to specialize it to deal with our tiles [98, 99].

To this aim, we exploited the Google's *Inception v3* model, made available for use as a pre-trained tool in the Keras library [100, 101].We used Inception as a feature extractor, with its weights (learned from the ImageNet dataset) frozen, by removing the classifying head and then replacing it with the new neural network to be trained. The model was trained for 10 epochs, as increasing the number of epochs yielded no significant difference when we came to the results. Not only, also class weights were employed (0.3 for non-site and 0.7 for sites, respectively), in an effort to help the model avoid mistakes on the positive cases.

Finally, in the hope of helping the model learn a more general representation, as well as to avoid overfitting, we also resorted to a data augmentation procedure that randomly applied a set of general geometrical transformations to all the images on a given batch. In this context we want to learn invariant representation to things like rotation and mirroring as they do not change the appearance of a site. Other popular transformation includes slight shifts in attributes like brightness, contrasts or in the color space. For this experiment, we resorted to the data augmentation procedures provided by the Keras library.

### 3.2.3 Results

First, we trained four models, combining two types of overlap rule (10% and 30%) to a different ratio of positive to negative examples (1:1 and 1:2). After that we tried two more variations: adding a larger context to the inputs and using a more aggressive data augmentation.

The AUC accuracy prediction scores for the hold-out test set (20% of the total) are shown in Table 3.1 and are mostly in the neighborhood of 0.60. Model 2 achieves the lowest at 0.55, while Model 3 is the best one, with a 0.63. These poor performances were somewhat expected given the difficulty of this task and lack of data. Nonetheless, we can conclude that a more conservative overlapping rule gives better performance overall and that the 1:1 ratio is consistently better, albeit slightly. Additional development led us to Model 5 and 6, documented in Section 3.2.5, showing an improvement with scores of 0.65 and 0.71 respectively.

Table 3.1: Classification performance for the 6 models we tested.

| Model | Type | Overlap | AUC Score (test) |
|---|---|---|---|
| 1 | 1:1 Ratio | 10% | 0.61 |
| 2 | 1:2 Ratio | 10% | 0.61 |
| 3 | 1:1 Ratio | 30% | 0.61 |
| 4 | 1:2 Ratio | 30% | 0.61 |
| 5 | Parallel | 30% | 0.61 |
| 6 | Aggressive Augmentation | 30% | 0.71 |

### 3.2.4 Using Prediction Heatmaps

Even if the general AUC metrics were not well promising, we decided to take the output from Model 3 and to overlay them to the geographical site map in order to print a pictorial impression of how the predictions came distributed over

a real map. In spite of the numerical results, in fact, our intent was to check if (at least visually) the predictions returned by the model could still point the user in the right direction, by highlighting a particular spot on the map. The following pictures (Figures 3.3 to 3.6) show an example area from the map and were produced by overlaying the predictions returned by Model 3 on the QADIS area map, using the software QGIS.

In particular, Figure 3.3 represents our starting point. The tiles in yellow should be predicted as sites, while those in blue as non-sites. Figure 3.4 should resemble a classical heatmap, where the more yellowish are the tiles, the higher is the associated probability they should represent (a part of) a true site, as predicted by the model (low probability zones are made transparent).

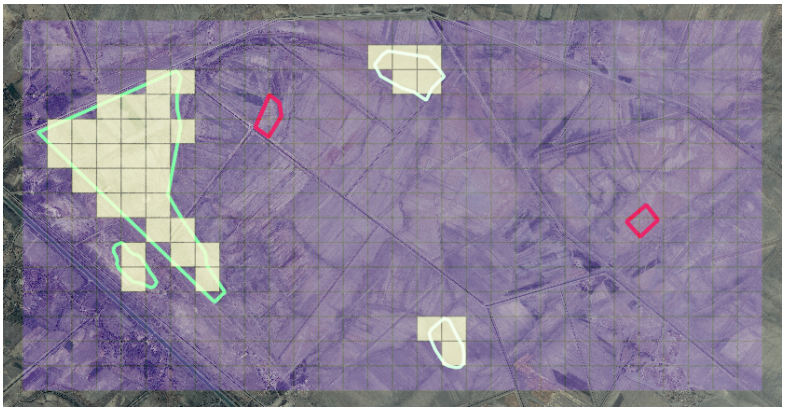


Figure 3.3: Ground Truth from the testing examples. Blue tiles are negatives, yellow are positives. Green contours are know sites, red are known non-sites.

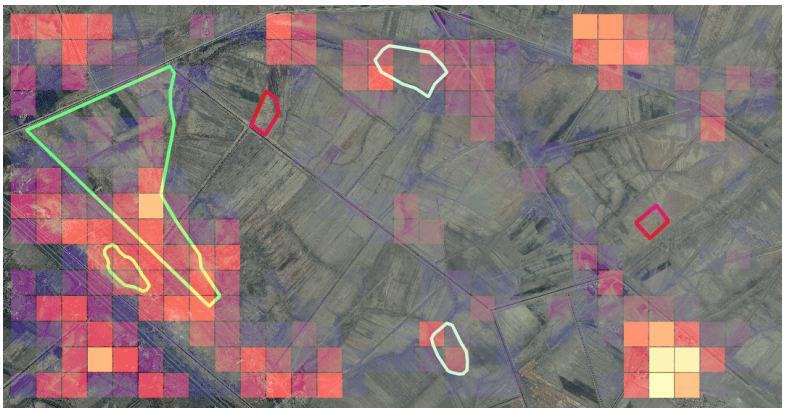


Figure 3.4: Prediction heatmap from the model. Dark colors corresponding to low scores are also made transparent for visualization sake.

While this result is still not convincing at all, we tried presenting the archaeologists a different visualization. The new idea we developed was to extend the prediction map to incorporate blocks of $5 \times 5$ tiles, over which the prediction probability was computed as equal to the maximum value of the contained tiles. With this new approach (Figure 3.5 ), the big archaeological site on the left is discovered, while the small sites on the right are still off our radar, and also two false positive areas, unfortunately, emerge. To notice is the fact that we adopted a very simple max function to reason with our heatmap. Less naive functions and filters could hopefully provide more accurate interpretations.

To check the robustness of our method, we tried to shift to the left all the scene, in order to verify if the predictions stay unchanged. Results are shown in Figure 3.6 and are controversial. On one side, it is confirmed that the big archaeological site is somewhat individuated (and the two little ones on the right stay off the radar), but more areas representing false positives emerge on the left of the scene, confirming that the approach is not stable.

Figure 3.5: Coarse-grained prediction of the same test example.

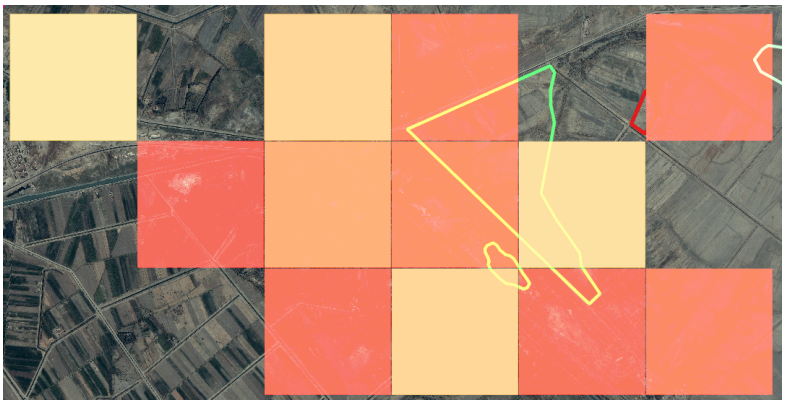


Figure 3.6: Coarse-grained prediction after shifting the input area to the right.

### 3.2.5 The Role of Context and Data Augmentation

Given the lackluster performance all the four models, we were suggested by the archaeologists that perhaps the inputs were too narrow to give the model a good sense of what a site is. To address this, we tried introducing a second input that would represent the surrounding context [102]. For each training tile we assembled a new image, composed of the 8 surrounding tiles around it, thus resulting in a $3 \times 3$ squared picture. This picture was then resized to a $299 \times 299$ tile, in order to feed it to our Inception v3 model (the same used to analyze the primary tile). As seen from Figure 3.7, we deployed a new custom model composed of two parallel Inception v3 branches, one for the tile and one for its $3 \times 3$ context, whose outputs were concatenated and fed to a dense layer for final prediction.

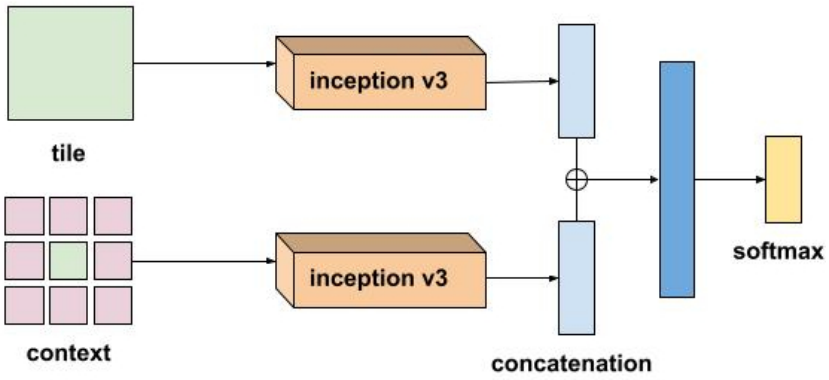



Figure 3.7: Parallel Architecture used for model 5 and 6. The idea is encoding a larger area using the existing tiles, constructing a 3x3 square and scaling it down.

The model was trained using the same hyperparameters of the Model 3: 30% overlapping; 1:1 ratio, 10 epochs; 0.3 / 0.7 class weights. After training, the testing activity was performed again, and a 65% AUC-ROC value was achieved, with just a moderate improvement in the prediction performances.

A final improvement to our neural model was achieved by carrying out a more intense activity of data augmentation [103]. While the first models used a randomly applied transformation at every iteration, this time we tried creating a dataset containing all possible symmetries of the input images (rotations and mirroring) to which then are randomly applied "destructive transformation" like shearing and zoom. This resulted into a quantity of tiles, representing positive examples (i.e., true sites), approximately equal to $12,776$. At that point, to work with a balanced training dataset (where the number of positive examples equals the quantity of negative ones), we further added $12,776$ tiles representing non-sites, and then we retrained our model.

The result we got, in terms of AUC-ROC on each single tile, was around the value of 71%. While one could acknowledge this as an important improvement, an increased performance in the classification of single tiles does not necessarily correspond to a better performance in recognizing true archaeological sites in their entirety, owing to the motivation that our tiles are, on the map, just small portions of larger archaeological sites.

## 3.3 Semantic Segmentation Approach for the *FloodPlains* Project

In this section we will look at a different approach to the same problem. The collaboration described in the previous section halted due the pandemic and once it was possible to continue, we had a bigger dataset available, coming from the *FloodPlains* Project [1].

This project has been developed in the framework of the European Union project *"EDUU – Education and Cultural Heritage Enhancement for Social Cohesion in Iraq"* [2], coordinated by Nicolò Marchetti. The ongoing project *"KALAM. Analysis, protection, and development of archaeological landscapes in Iraq and Uzbekistan through ICTs and community-based approaches"*, funded by the Volkswagen Foundation has allowed a review of our data input and the development of the research presented in this section[3]. The CRANE 2.0 project of the University of Toronto provided the geospatial servers on which FloodPlains is running.

Given the size of this new dataset, just shy of $5,000$ examples, and the limitations of the previous approach (namely, the cumbersome tiling process and the impossibility of working with images of arbitrary size) we decided to frame the problem as a semantic segmentation task this time, and use pretrained deep learning models to achieve the best performance. This choice allows us to overcome the limitation in input size of the previous approach and generates pixel-level classification maps that are way more useful for indicating the presence of sites. Using this type of machine learning approach is not unseen in archaeology but most works focus on the use of Random Forest models trained on very small and narrowly focused datasets [104].

### 3.3.1 Dataset

We started with a dataset of geo-referenced vector shapes corresponding to contours of known Tell sites in the survey area of the Floodplains Project that spans more than $66,000 km^2$, as shown in Figure 3.8. This dataset contains $4,934$ shapes collected from a variety of sources, dating back even several decades, and who were all confirmed by in person surveys.

To generate a dataset of images we imported the shapes mentioned above into QGIS (an open-source GIS software) and using a Python script we saved a square of length L centered on the centroid of site which contains only satellite imagery

[1] https://floodplains.orientlab.net
[2] EuropeAid CSOLA/2016/382-631 `www.eduu.unibo.it`
[3] www.kalam.unibo.it

from Bing Maps (we also considered other maps but found that in this particular area they all are mostly the same). After this we saved the same image without a base map but with the site contour shown, represented as a shape filled with a solid color, to serve as our ground truth masks.

In the first experiments we set $L$ to be 1000 meters but, after consulting with domain experts we imagined that the increasing the size of the prediction area could be beneficial due to the inclusion of a larger context. Subsequently we also tried using $L = 2000$ meters with improved results.

From this square we randomly crop a square of length $L/2$ to be used as the input. This ensures that the model does not learn a biased representation for which sites always appear at the center of the input, Figure 3.10a shows, and additionally serves as data augmentation. When extracting from QGIS we saved images with a resolution of around 1 pixel/meter (1024 pixels for 1000 meters, double that for the model with increased input size) but the inputs were then scaled down to half of that to ease computational requirements while having low impact on the overall performance. We split the dataset into a 90% training set and a 10% holdout test set, stratifying the "empty" images we added.

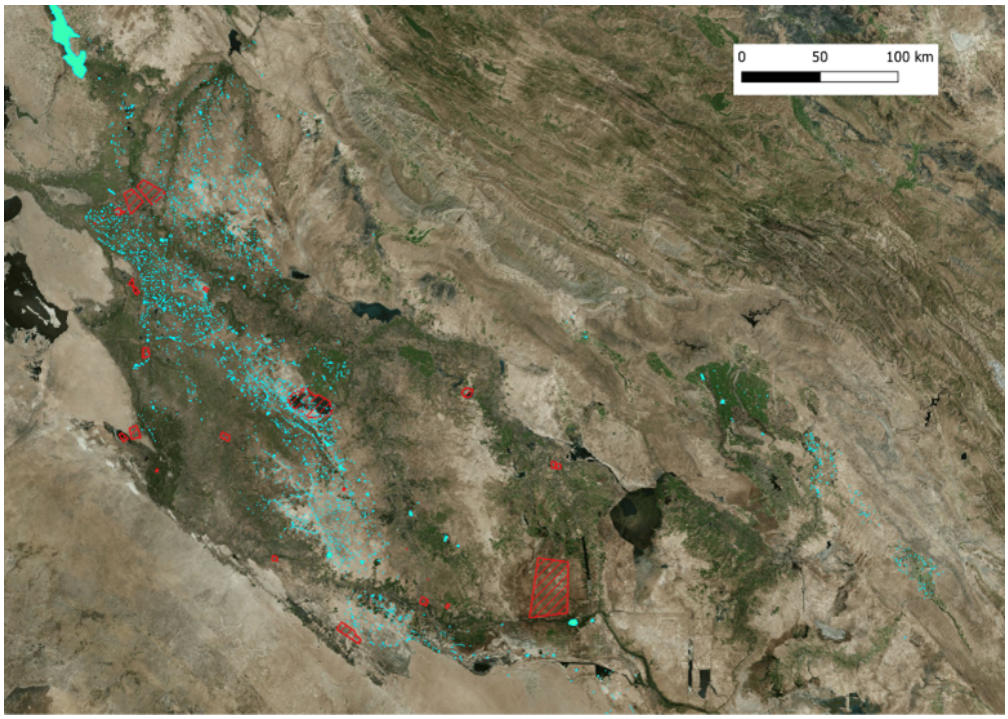



Figure 3.8: Investigation area. Cyan shapes represent surveyed sites. Red areas are location where no site can be found, like cities and artificial lakes.

Following a suggestion from domain experts, we also tried integrating *CORONA* imagery as an additional input [105] as in their usual workflow those historical aerial photos are very useful and often combined with the satellite base map and other topographical maps. After importing the photos into QGIS we followed the same procedure to create the inputs, ensuring the crop operation was equal for both Bing and Corona images.

### Filtering out bad examples

After inspecting the outputs of the first experiments we noticed some site shapes looked wrong, as shown in Figure 3.9. The archaeologists then warned us that the dataset was compiled as a comprehensive source of information for their workflow, rather than specifically to train a machine learning model, and highlighted that some of the shape could either be imprecise or not visible on present day photos (but is in historical ones). To gauge the effect of this impurities we also wanted to try training models with a filtered dataset. Theoretically those sites provide no information and could actually impair the learning process. Exploring the dataset, we found that some sites that were too either too big or too little, as well as sites that are not visible anymore from present day images because of agriculture, urbanization, or flooding

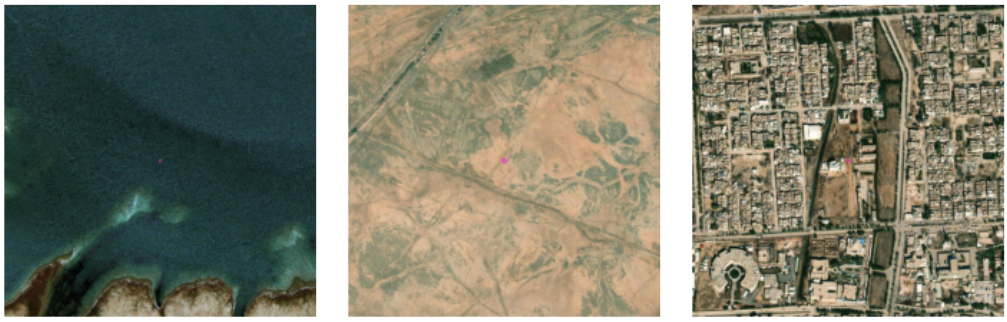


Figure 3.9: the examples from the filtered images. The sites are either flooded, extremely small or covered by a city.

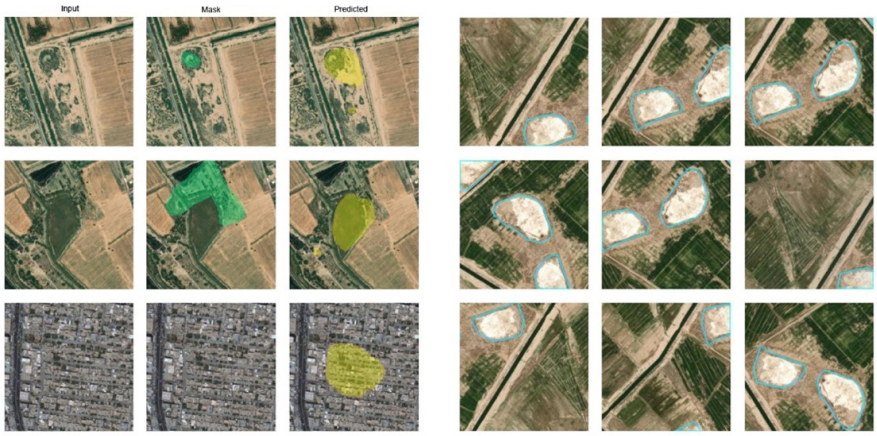
For the filtered dataset, we started by removing the top 200 sites by area as these were considerably bigger than the rest of the dataset. Visual inspection confirmed that they follow the shape of areas that are not just simply Tells. After a discussion between data scientists and archaeologists we convened that this was a good heuristic as it helped focusing the target only to tell shapes. Besides, such large areas would be difficult to learn and arguably meaningless, as the mask to be predicted would be completely filled with ones. The number 200 comes from calculating those that would not fit inside the squares we used as inputs

and translates to excluding sites with areas larger than 20 square degrees (around 0.25 $km^2$).

Additionally, we filtered out 684 sites that either presented a very small area or contained notes by the archaeologists that suggested they were not visible anymore. In particular, the size threshold was set at 0.1 degrees squared (roughly equal to 1000 $m^2$). This very small sites actually correspond to a generic annotation for a known sites with unknown size or precise location. After this procedure, the total number of shapes is thus 4050.

### Data Augmentation

As in the tile-based classification model, we performed data augmentation and leveraged the Python library *albumentations*, which provides a framework for easily applying a variety of transformation with a random probability at load time [106].




(a) (b)

Figure 3.10: a) prediction from the model trained with no random cropping (ground truth in green, prediction in yellow); b) Nine examples of possible augmentations for the same site (cyan contour).

Apart from the random crop we described before, which was always performed, we included three types of augmentations. They are as follows: The first is a random rotation of 90, 180 or 270 degrees. This transformation is non-destructive, and it is useful to teach the model that an archaeological site is recognizable, regardless of the orientation. Similarly, we also applied a random mirroring, either horizontal or vertical covering all the possible symmetries an image can have. Lastly, we applied a (slight) brightness and contrast shift, as the images are not uniform in their lightning conditions, and this should help the model recognize sites in those cases. Figure 3.10b shows some examples.

As a final addition, we introduced 1,155 images with empty masks (no sites to predict) sampled from locations suggested by the archaeologists (red areas in Figure 3.8). Those include highly urbanized areas, intensive agricultural areas, location subject to flooding (i.e., artificial lakes and basins) and rocky hills and mountains. The total number was arbitrary and chosen by considering the size of each suggested area and of the tiles. This addition brought the final number of images to 5,025.

#### Uzbekistan

At the request of the archaeologists, we also performed an additional test on a dataset coming from the *Uzbek-Italian Archaeological Project* [107]. Given the similarity between the Tell in the Mesopotamian floodplain and the Uzbek Tepa in the we wanted to see if the model was able to detects those sites without the need of additional retraining.

The dataset features 2,318 point-like annotations, categorized in different ways, which also come with attributes related to their preservation states. Following the indication of domain experts, we selected only sites classified as either *Tepa*, *Low Mound* or *Monument* with the well-preserved label. The final number of sites ends being 229: 148 Tepa, 67 Mounds and 14 Monuments. The actual test set images were created following the same procedure described previously.

### 3.3.2 Models

All deep learning models for semantic segmentation are based on the same architectural concept of employing both an encoder and a decoder. The encoder is in charge of feature extraction, and at various levels of detail, of reducing the image to smaller and smaller feature maps, essentially learning where to look in the image. The decoder instead plays the role of inflating the feature maps back to the input size, while learning to create the actual mask one wants to predict.

We employed a library of pre-trained segmentation models for Pytorch, as the primary goal of this study was to check its feasibility [108]. The library in fact allows the use of different segmentation architectures, that in turn shape the decoder section, and combine them with different encoders for feature extraction.

In our experiments, we used *U-Net* and *MAnet* as the segmentation architectures, and *ResNet* and *EfficientNet* as encoders for feature extraction. U-net is a fully convolutional network architecture introduced in 2015 for semantic segmentation of cellular tissues. The model is characterized by two almost symmetrical halves, an encoder and a decoder, hence the U shape that gives the name, with connections that go across at the same depth level [109]. MAnet was also devel-

oped in the context of medical applications but it also showed successful results in segmentation tasks involving satellite images [110, 111]. It features attention blocks (made popular by the transformers architecture) and an architectural design aimed at better capturing long range spatial dependencies ResNet is a very popular and influential deep learning architecture for computer vision introduced in 2015. It popularized the idea of skip connections, becoming the state of the art for convolutional neural networks (CNN), and it is often used as a benchmark for new models. We employed a version with 11 million parameters (i.e., resnet18), pre-trained on the ImageNet dataset [112]. EfficientNet is an optimized convolutional network introduced by Google Brain, that features a streamlined architecture thanks to clever design decisions and to the use of a neural architecture search to find the best scaling for depth, width and resolution. We used the B3 model which has a similar amount of parameter to resnet18, while allegedly performing way better [113].

Finally, some words are in order regarding loss functions. Loss functions play a very important role as they are directly responsible for the way the model learns and thus produces its outputs. Among the many alternatives, we used the Intersection-over-Union (IoU) metric. This serves as our performance metric while the formula below shows how it is computed:

### Metrics and Loss Functions

The metric of interest in semantic segmentation is called *Intersection-over-Union* (IoU). As the name implies is measure the ratio between the intersection and the union of the predicted mask and the ground truth mask. The formula below shows how it is computed in terms of the outcomes of pixel classification:

$$IoU = \frac{Y \cap \hat{Y}}{Y \cup \hat{Y}} = \frac{TP}{(TP + FP + FN)} \tag{3.1}$$

where $Y$ is the segmentation mask and $\hat{Y}$ is the predicted mask (TP, FP and FN stand for true positives, false positives, and false negatives).

While IoU can be used a loss function, for numerical differentiation reasons Dice Loss is often preferred [114].

$$L_{dice} = 1 - \frac{2TP}{2TP + FP + FN} \tag{3.2}$$

We also experimented with Focal Loss. This is a variation of the classical Cross Entropy Loss with the introduction of a mechanism that scales down the

contribution of easy to predict elements [115]. For each pixel to classify we have:

$$L_{focal} = -\alpha(1 - p_t)^{\gamma} log(p_t) \tag{3.3}$$

with

$$p_t = p \cdot \text{target} + (1 - p) \cdot (1 - \text{target})$$

where *target* corresponds to each pixel in the mask (either 1 or 0) and p to the predicted probability.

### 3.3.3 Results

The first experiment we performed were aimed at exploring if the use of different models made a difference in the final result. all models performed quite well, with testing IoU scores around the value of 0.70, as summarized in Table 3.2. Not shown are the foreseeable exceptions of the base model (U-net with resnet18 and dice loss) trained either without cropping and negatives, or with only cropping which scores respectively around 0.50 and 0.64.

Table 3.2: Validation Performance (per-image IOU).

| Model | Architecture | Encoder | Loss | Epochs | IoU |
|---|---|---|---|---|---|
| 1 | U-net | resnet18 | dice | 10 | 0.6823 |
| 2 | U-net | efficientnet-b3 | dice | 10 | 0.7068 |
| 3 | U-net | resnet18 | focal | 10 | 0.7221 |
| 4 | U-net | efficientnet-b3 | focal | 10 | 0.7219 |
| 6 | MAnet | efficientnet-b3 | dice | 10 | 0.6920 |
| 7 | MAnet | efficientnet-b3 | focal | 10 | 0.7265 |
| 8 | U-net | efficientnet-b3 | dice | 20 | 0.7178 |
| 9 | MAnet | efficientnet-b3 | dice | 20 | 0.7316 |
| 10 | MAnet | efficientnet-b3 | focal | 20 | 0.7437 |
| 11 | MAnet (filtered dataset) | efficientnet-b3 | dice | 20 | 0.7662 |
| 12 | MAnet (2k input) | efficientnet-b3 | dice | 20 | 0.8076 |
| 13 | MAnet (2k input+filter) | efficientnet-b3 | dice | 20 | 0.8154 |
| 14 | MAnet (1k input+CORONA) | efficientnet-b3 | dice | 20 | 0.7406 |
| 15 | MAnet (2k input+CORONA) | efficientnet-b3 | dice | 20 | 0.8345 |

MAnet seems to provide no significant benefit over U-net, at least in our experiments: their predictions are extremely similar in most cases and so are their scores. MAnet seems to be taking the lead slowly with more training iterations, though. Similarly, Dice Loss and Focal Loss obtain extremely close scores, with the only discernible difference being the output they produce. Dice Loss tends to create masks that are more cohesive and clear-cut, with blocks of high probability that sharply taper off to 0.0 at the edges, whereas Focal Loss creates hazier prediction maps that change more smoothly. Coming to the encoder choice, as expected resnet18 consistently performed worse than efficientnet-b3. However, the gap is not extremely marked in terms of IoU, even though qualitatively it seems to make worse mistakes (e.g., it misses some part of the sites).

We decided to go further with our experiments by only using the combination of MAnet with efficientnet-b3 and focal loss, trained for 20 iterations. We retrained a model with these settings on the filtered dataset which resulted in an even improved IoU score. This is likely due both the effect of better learning and less "false" mistakes in the scoring process.

Finally, we tried to double the size of the inputs after noticing with the domain experts that the model seemed to make mistakes in situations where there was not enough context to understand the target. This resulted in a considerable increase in performance, reaching 0.8076 and going up to 0.8154 when the filtered dataset is used.

The inclusion of CORONA images in the input did not seem to meaningfully change the performance of the model with the smaller input size, with an IoU of 0.7406. This could be likely due the low resolution of these images. On the other hand, the model with the larger input received a boost in segmentation performance, scoring 0.8345 IoU, although a quick look at the prediction showed that there was not a striking difference from the model without CORONA inputs. As we will see next, testing for site detection showed that the model is not actually performing better.

### Detection Performance

To measure the detection performance, we transformed the raster predictions from the best models into vector shapes, using the well know library *GDAL* (Geospatial Data Abstraction Library), and looked for the intersection between the site annotations and the predictions.

To obtain smoother shapes, before the conversion we first applied a Gaussian blur to the prediction raster and then clipped values above a certain threshold (0.5 but the number can be changed for a more or less sensitive model) to 1.0 while

everything else would be set to 0.0. This automatic evaluation gives good but not too exciting results, with an accuracy score of 62.50% and 60.08% for model 13 and 15 respectively.

We can argue already that the would provide a good starting point for human analysis, being able to find most sites reliably. However, given the unreliable nature of the dataset, we wanted to involve the domain experts into the process to provide a verification of the predictions and to differentiate the cases in which the model commits proper mistakes from those in which it makes justifiable errors that a human would do too [116].

Table 3.3 summarizes the results of the automatic and the human evaluation, showing a marked improvement in detection performance given by re-adjusting mistakes according to their nature. We include recall, besides accuracy, as the percentage of real sites retrieved by the model is perhaps the most important aspect for the archaeologists.

Table 3.3: Site detection performance for the best model: automatic and adjusted by human review.

| Model | Evaluation | TP | TN | FP | FN | Accuracy | Recall |
|---|---|---|---|---|---|---|---|
| 13 | Automatic | 228 | 98 | 70 | 125 | 0.6257 | 0.6459 |
| | Adjusted | 258 | 185 | 40 | 68 | 0.8040 | 0.7914 |
| 15 | Automatic | 209 | 104 | 57 | 151 | 0.6008 | 0.5806 |
| | Adjusted | 239 | 197 | 27 | 88 | 0.7913 | 0.7309 |

Let us explain how this adjustment procedure was carried out. First of all there a considerable number of sites that are no longer visible from present day satellite images and were not filtered from the dataset. This was expected as only 50% of the annotations had additional information and even less contained indication of their visibility. Those sites should not be considered as False Negatives but rather as True Negatives. When it comes to predictions marked as False Positive, sometimes the model predicts another site close by, instead of not the being tested. The can be considered a mistake or not depending on the nature of the "missed" site. In one case we have a site that is not longer visible, so the prediction is actually a True Positive. On the other hand, it can be a site that is still visible, but maybe less so than another one, close by in the picture. Given that, in a real world scenario, the closeness to other sites would result in a useful suggestion as the human expert would then be able to retrieve them all, we considered those sites as True Positives. We could alternatively consider them as both a false negative and a true positive, or even avoid considering non visible sites altogether resulting in a minimal difference with accuracy 0.7837 and recall 0.8201. Lastly,

some predictions were actually present in the outputs but too faint for the cutoff threshold we imposed. We did not adjust for those errors but they indicate a possible approach for interaction: using predictions as overlays and manually looking at the map. Alternatively setting a lower threshold could solve the problem.

It is interesting to see how Model 15 is performing worse, even if it had higher IoU. Looking at the predictions, it appears that this model is a little more "cautious", resulting in a lower recall but also less False Positives. In turn, this can result in a higher IoU because it reduces the Union term, and if areas are a little bit more precise it even raises the Intersection term. However, for detection's sake, we need the presence of an intersection rather than a perfect match and in this situation the lower number of positives is punishing. Overall the difference in accuracy is not excessive but we must also consider the additional complexity and cost of using two sets of input images which make model 15 a bit cumbersome. For this reason we moved on using just model 13.

To conclude, in Table 3.4 we show some examples of the comments provided by the archaeologist regarding the prediction produced by model 13, which are instead showed in Figure 3.11.

Table 3.4: Nine examples of comments from the archaeologists on the model predictions showed in Figure 3.11.

| Site | Comment |
|---|---|
| AKK.0006 | The site shows various levels of destruction due to modern structures, canals and fields. Difficult to spot on Bing images. |
| AKK.0021 | Site completely destroyed by agriculture. |
| AKK.0106 | Site completely destroyed by agriculture. |
| AKK.0213 | Correct prediction. The model also recognizes as a site a portion not included by the archaeologist but that is certainly part of it. |
| AKK.0261 | Correct prediction. The predicted site SE of the target, could effectively be labled by an archaeologist doing remote sensing. |
| AKK.0317 | Correct prediction. Although the target is not predicted because of its location, the marked sites in the NE would be also labeled by an archaeologist doing remote sensing. |
| AKK.0355 | Perimeter is very faint and is difficult to locate from Bing images because of agricultural destruction. |
| AKK.0621 | Correct prediction. Site in NW of the target would be also labeled by an archaeologist doing remote sensing. |
| AKK.0776 | Site is completely destroyed by urbanization. |

(a) AKK.0006 (b) AKK.0021 (c) AKK.0106

(d) AKK.0213 (e) AKK.0261 (f) AKK.0317

(g) AKK.0355 (h) AKK.0621 (i) AKK.0776

Figure 3.11: Nine example predictions as mentioned in Table 3.4. Site outline is shown in Green. Yellow areas are True Positives, Orange areas are False Positives.

### Maysan

After that we also tried the model on rectangular area from the Maysan region for which we obtained annotations that we did not include in the initial dataset. This test had the goal of evaluating how many false positives the model would predict and to give an example of the mistakes the model makes in an operational scenario.

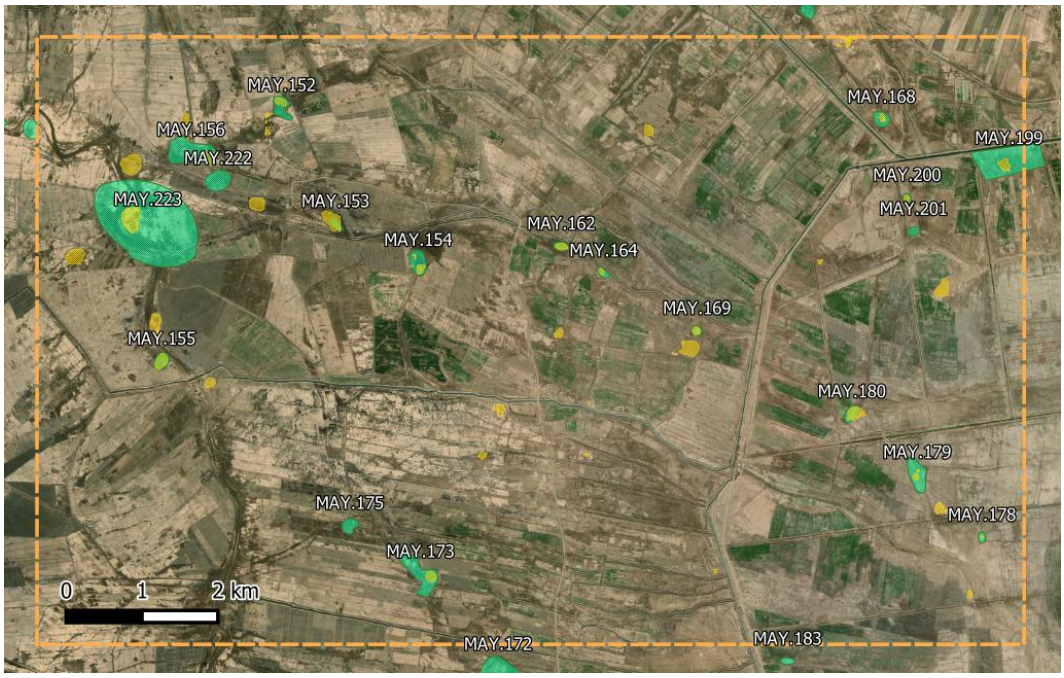



Figure 3.12: Maysan test area (orange) with ground truth sites (turquoise) and predictions (yellow).

The area we selected contains 20 sites and span 104 $km^2$. Figure 3.12 shows the area with the annotation from the archaeologists and the prediction from the model. As it can be seen, the model is able to recover 17 of the 20 sites while also suggesting around 20 more shapes (depending on what is considered a single instance). Most of these suggestions are not useful but are also easily and quickly sifted out by an expert eye, especially in context, given their size or their location.

### Uzbekistan

On the Uzbek dataset the situation is not as good unfortunately. Evaluation of the outputs showed that the model is able to identify correctly around 25 to 30% of the sites in this region, depending on how strictly the shapes are considered.

The errors are comprised either of sites that are missed completely or sites that are somehow hinted either too faintly or inside a huge area that appears meaningless.

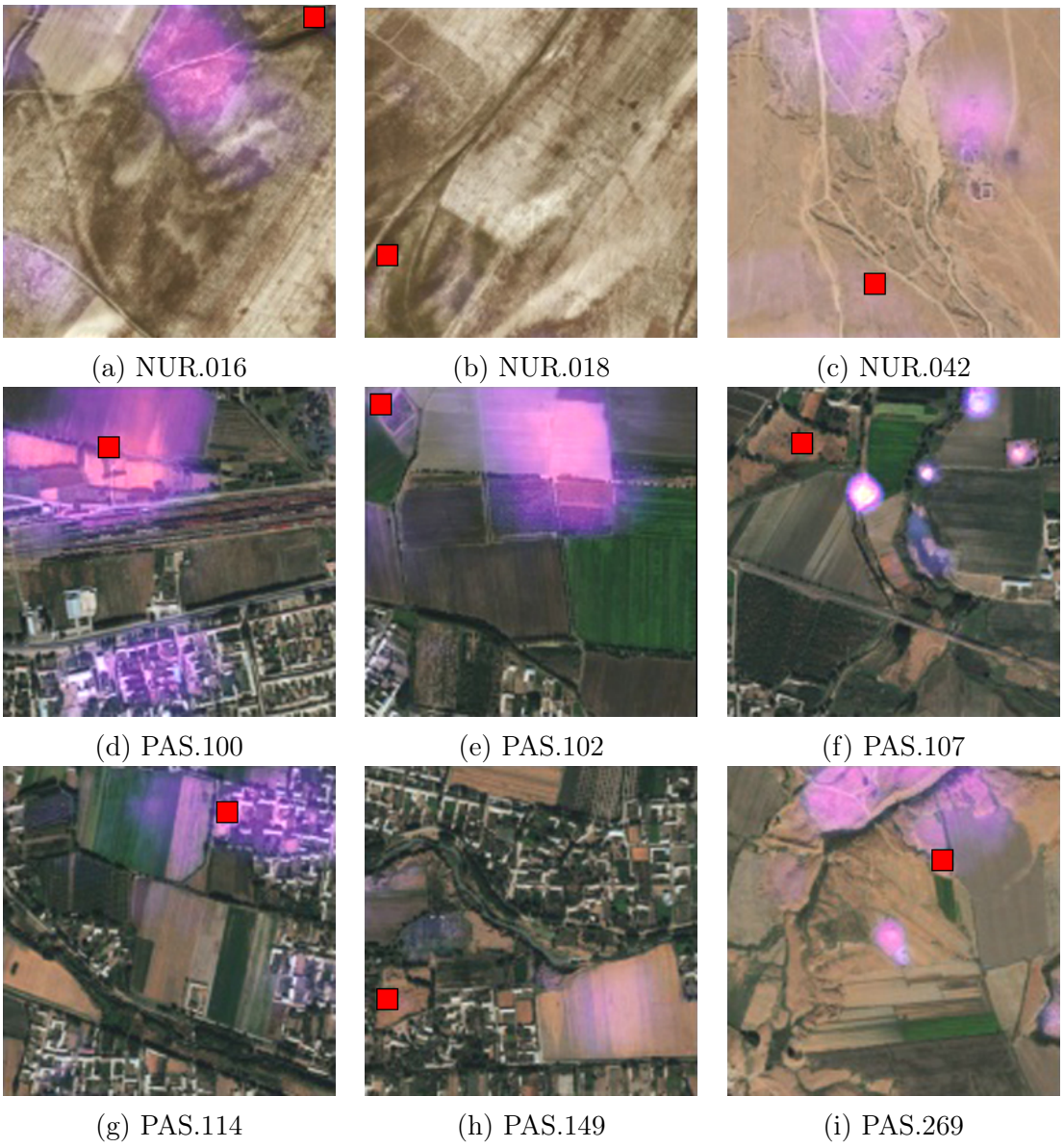
The reason for this severe drop in performance is most probably due to the different nature of the landscape in the region which in some locations appear to be way more urbanized and in general features more vegetation. Furthermore, the conventions which lie behind the annotations in the Uzbek dataset might not be perfectly aligned with the Mesopotamian one further complicating the situation.

The only way of dealing with this problem is probably to create a small dataset of selected Tepa sites and perform an additional round of transfer learning so that the model may grasp the new context and characteristics in the region.

Table 3.5 contains some comments the archaeologists provided after looking at the prediction on the Uzbek dataset. The sites mentioned here, are shown in Figure 3.13 where the annotated sites are shown with a red square and the prediction are ovelaid on the image as we did not perform any automated test in this case.

Table 3.5: Nine examples of comments from the archaeologists on the model predictions showed in Figure 3.11.

| Site | Comment |
|---|---|
| NUR.016 | Difficult to locate as inside deep paleorivers that make the landscape difficult to read (experts tend to focus more near those physical features). |
| NUR.018 | Very small site difficult to locate in this landscape. No chromatic variation and faint elevation. |
| NUR.042 | Very small site difficult to locate in this landscape. Many traces of riverbeds make the landscape difficult to read. |
| PAS.100 | Small Tell (14m diameter) easily recognizable among farmed land but probably too small compared to the training data in Iraq. |
| PAS.102 | Similar to PAS.100, too small compared to Iraqi tells. |
| PAS.107 | Correct prediction. The predicted areas would be considered by an archaeologist doing remote sensing. |
| PAS.114 | Tell is visible but surrounded by densely urbanized landscape. |
| PAS.149 | This tell does not have the traditional elliptic shape as it was partially salvaged from urbanization. |
| PAS.269 | Landscape full of potential tells, the predicted area should be surveyed. |



(a) NUR.016 (b) NUR.018 (c) NUR.042

(d) PAS.100 (e) PAS.102 (f) PAS.107

(g) PAS.114 (h) PAS.149 (i) PAS.269

Figure 3.13: Nine example predictions as mentioned in Table 3.4. Site outline is shown in Green. Yellow areas are True Positives, Orange areas are False Positives.

### 3.3.4 Proposing a Human-AI Collaboration Workflow

With a model such as the one we just described we can imagine a new workflow for the archaeologists doing remote sensing.

Starting from the dataset the model produces prediction masks that can be manipulated through post-processing to obtain a vector shapefile that can be used for automatic evaluation and detection of sites. At this stage, the user has the possibility of choosing how a threshold to cut prediction off and the use of techniques to smooth the output shapes like blurring or buffering the vectors. All these operations can be done automatically by the model code or left to the user which can employ common tools in any GIS software, where is also possible to choose the preferred graphical representation of the outputs.

Additionally, a map overlay can be generated by stitching together adjacent prediction maps and visualizing the probability values, resulting in something similar to an heatmap. The goal in this case is that of spotting sites that might pass undetected by the automatic comparison because their probability lower than the threshold while still being distinguishable for humans who are also able to integrate their intuition and contextual clues coming from other maps and sources of information.

Each time model is used, in either way, after reviewing the outputs the users would be able to obtain either a new set of annotations or a list of sites to be removed or relabeled. Figure 3.14 summarizes the use we envision for the model we described.

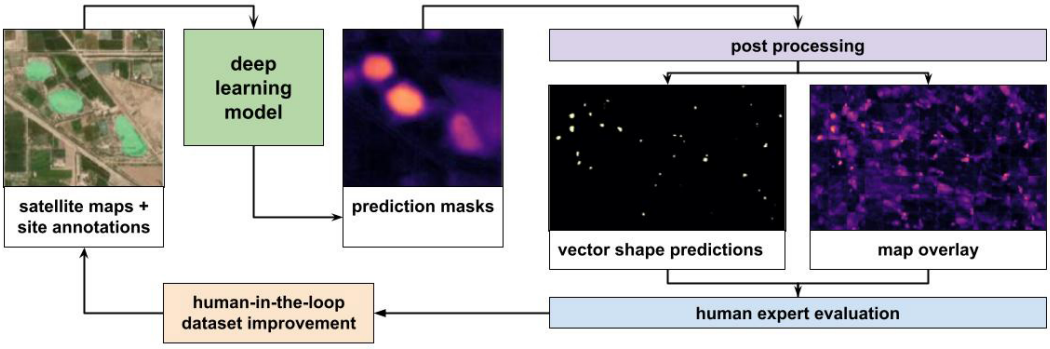



Figure 3.14: Proposed Human-AI collaboration workflow

We demonstrate the overlay tactic on a small area in the Maysan region, as shown in Figure 3.15 but the computation could be easily scaled up to cover huge areas, as it takes less than a second to produce an output and there is no need to complete the operation in one go anyway.

The only shortcoming of this method is the evident ridge between different input images. In theory semantic segmentation could work with inputs of arbitrary size but doing so requires huge amount of memory which might not be available. A solution might be the creation of overlapping prediction maps that would then be averaged, trading off computational time for increased precision.

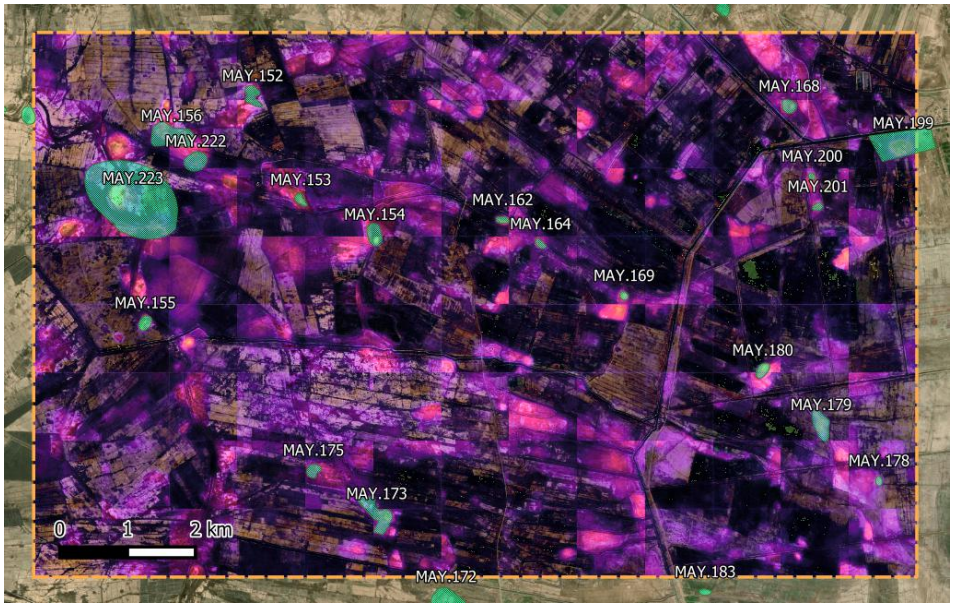



Figure 3.15: Map overlay example in small portion of the Maysan region

## 3.4 Conclusion

This chapter presented a challenging task where the interaction with the human experts is very important and goes way beyond the optimization of scoring metrics. The conclusion we can draw from this experience related to Research Questions 1, 2, 3 and 4.

Answering **RQ1**, we saw how important the expertise of archaeologists has been in this project. Together we were able to define the dataset and decide how to treat the annotations representing the site shapes, trying to establish what counts as a positive example and what counts as a negative example. Their input helped us move from the classification to the segmentation approach which demonstrated more effective. Moreover, they helped us filter out of the dataset all the instances that could throw off the model in its learning process.

Concerning **RQ3**, we saw how human evaluation was necessary as in this context an automated process fails to show if a prediction is actually good or bad on its own. IoU can be low when model has predicted the correct site in a sub-optimal way or when it has predicted the site perfectly but also something else that is not useful (but that the humans could easily ignore). Additionally, we can have a low score for a training example that is not visible anymore that the model correctly ignores, resulting in punishing the right behavior by the neural network. This consideration carry over when we try to measure detection performance by considering intersection between the segmentation and the sites. A proper evaluation then, should try to understand where the model is making justifiable mistakes that can either be:

- Mistakes that can be explained by lack of visible sites, which a human would make too
- Mistakes where the model suggests a probable site that it is not yet in the dataset, but that human would also consider
- Predictions which point in the right direction even if the shape is not the same as the one in the dataset

Using human-in-the-loop evaluation and adjusting the scores on the test set shows the model is more apt than it appears at first glance. Even when considering detection performance, we see how metrics are insufficient to capture the performance and utility of our models. Once again, not every mistake has the same weight in this contexts and we could argue that a more sensitive model (i.e., with higher recall) is more useful here. Furthermore, addressing **RQ2**, our colleagues in archaeology told us that the mistakes our systems make are not problematic for their use case and that instead the speed and simplicity is remarkable. The

usefulness of an artifact such as our deep learning models highly depends on the problem at hand and the way it was addressed before, without the use of intelligent systems. Even an under-performing system can still yield improvements if not in terms of accuracy perhaps in terms of time.

We concluded the chapter by proposing a human-AI collaboration workflow that allows the domain experts to benefit from the model, even as it is imperfect, in term of time savings. In turn they can make it better by correcting its mistakes along the way, while doing the work they would normally do. One of the interaction modes we discuss involves the use of the model to produce an overlay visualization instead of just giving predictions. Considering **RQ4**, This overlay can syngergize with the knowledge of the archaeologists and focus their eye on interesting areas even if they would not be considered as a site by the automated procedure, improving the overall results.

# Chapter 4

# Symbolic Music Generation with Transformers

In this chapter we will describe the results of a project focused on the generation of symbolic music. Music is a peculiar field of application when compared to the previous two we talked about because, being an art form, it has clear rules and structure that can be learned but it lacks an easily identifiable metric to define its quality [117]. There are certainly many ways to judge and critique music but none of those is more valid than the other and they do not require to be universal and perfectly coherent. This in turn makes it difficult to formalize a set of rules or a function to score some piece of music according to any concept of quality.

Modeling music can be achieved using the same techniques used for generating text. In fact, we can think of symbolic music and its relationship to music itself in the same way we think about text and its relationship to language. The only difference is that for the musical style that include polyphony, there is also a vertical relationship between instruments or voices, while natural language is mainly horizontal with relationships spanning different words in a text.

The two main approaches to generating sequences of notes, or words, are those based on rules and those based on statistical modeling [118, 119]. The latter is the current state of the art, as the former proved to be not flexible enough to adapt to the many different forms language and music appear in the real world, while also requiring a great deal of effort in order to formalize a "grammar". Statistical modeling can be reduced to performing the task of predicting the next word or note in a sequence. The various applications that leveraged neural networks in the last few years showed us that this is enough to learn structures and rules from a dataset of meaningful examples without the explicit coding of rules [120].

Both music and text generation however face the issue of evaluating more abstract concepts like style or any type of aesthetic value. In simple words: how do we defined what a good piece of music is? What makes a compelling text or poem? The only way to answer these answers is with the involvement of human experts and users in the creation of the dataset, in the development of the models and in the evaluation of the outputs.

These themes are a central part of the MUSAiC project, in the context of which this project was carried out [121].

The object of our efforts was initially Irish folk music and the implementation of a new neural network based on transformers that could surpass the performance of folkRNN the previous state of the art in this task [122]. Achieving this goal proved to be more complex than expected and highlighted the importance of human-in-the-loop techniques in order to solve issues during development. In particular, hyperparameter tuning was hard to perform by simply looking at the loss curves and required human expertise and manual evaluation of the outputs. In combination with informative data visualization this allowed us to guide the design decisions and obtain the performance we desired after designing a new sampling strategy that could compensate the weak points of the model.

The final confirmation of the quality of the work came from the participation in the AI Music Generation Challenge 2021.[1] This competition focused on a style of Swedish folk music called slängpolska which is quite different from Irish folk. The Tradformer proved to be flexible enough to learn the style from the few examples available by leveraging the knowledge it got from Irish music and it received the highest scores from the judges, winning multiple prizes.

The Chapter is structured as follows: Section 4.1 discusses data representation and the datasets we used, both Irish and Swedish; Section 4.2 describes the deep learning models involved and the visualization techniques we employed; Section 4.3 explains the difficulties we faced during development and illustrates the human-in-the-loop evaluation; Section 4.4 deals with the Swedish music task and how we performed transfer learning. Section 4.5 is a brief discussion of how the Tradformer was used in music co-creation sessions and how generative models relate to creativity; Finally, Section 4.6 concludes the chapter.

[1] `https://github.com/boblsturm/aimusicgenerationchallenge2021`

## 4.1 Dataset and Representation

Let's start by detailing the dataset we used and the peculiar musical representation it features.

### 4.1.1 The `abc-notation` Standard

We employed a particular type of symbolic format called `abc-notation`. This standard has a long history in world of Irish folk music and has seen a wide adoption by online communities of musicians as it easily adapted to the new communication medium that was the internet. It was so popular that it has a specific MIME type.[2]

Let us go over the most important tokens in the abc standard that we decided to include in our model's vocabulary. It is worth noting that while we were interested only in the most basic elements that pertain to the melodic content, the notation supports the vast majority of elements that could be found on a sheet music, even for more complex musical traditions. However, since this project focused on Irish and Swedish folk music which is mostly monodic and where expression and embellishment are usually left to the players, learning any indication beyond the actual notes in the melody was not important.

The abc format uses letter A to G to indicate pitches following the Anglo-Saxon convention. Natural numbers after each letter indicate the duration of that note with respect to the basic step (usually 1/8). Fractions can be used to indicate a duration shorter than the base step or the write notes that do not have a integer multiple duration (e.g a dotted semiquaver can be expressed using `3/2`) The most basic version covers two octaves using uppercase and lowercase letters, but this range can be extended by adding one or more commas or a ticks after the letter to indicate lower or higher octaves (e.g. `C,` or `c'`). Alterations use the symbols `^` for sharps, `_` for flats and `=` for naturals.

Beside note and duration there a number of "structural" tokens to be considered: Each bar is delimited by the `|` token, while `||` signals the end of a tune. Similarly reminiscent of sheet music notation we also have and `|:` and `:|` signifying the start and end of a repeated section. `|1` and `|2` are used to indicate a first and second ending bar, which are quite common. The occasional triples is indicated by `(3`, and chords/double stops are contained in squared brackets `[ ]`.

Finally, each tune begins with a series of fields that represent metadata: `X` indicates the id of that tune in a collection; `T` is for title; `C` for the composer; `M` specifies the meter and `K` the key and modality (e.g. C dorian); `L` is for the base step duration and `Q` for the base metronome indication.

[2] `https://www.iana.org/assignments/media-types/text/vnd.abc`

Figure 4.1 shows an example of this notation and Figure 4.2 shows the corresponding rendered version on the staff.

```
X:754
T:Slangpolska fran Barseback
T:(SvL Skane nr. 754)
T:efter Per Munkberg
O:Barseback, Skane
R:Polska
M:3/4
L:1/16
K:Am
|:e4A4E4|e4 B2c2 e4|cdef agfe cdef|gfed fedc B2A2|e4A4E4|
e4 B2c2 e4|cdef agfe cdef|gfed f2e2 e4:|
|:GBeB GBeB GBeB|FAeA FAeA FAeA|GBeB GBeB GBeB|
FAeA FAeA FAeA|ABcd e4 a4|gfed f2e2 e4:|
```

Figure 4.1: abc-notation for the tune Slängpolska från Barseback

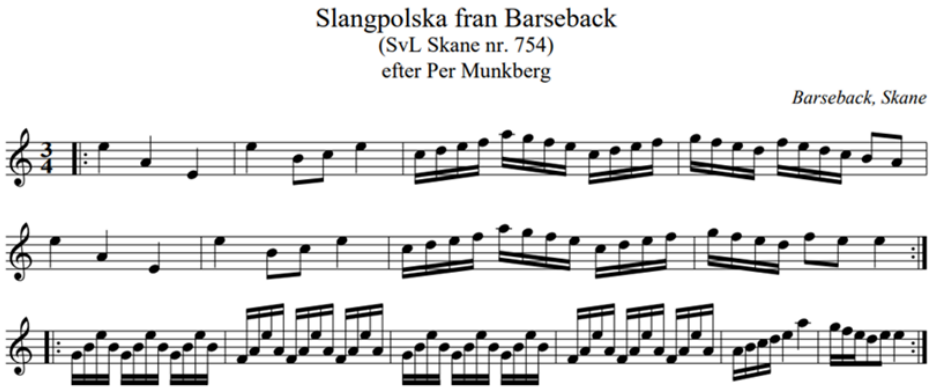



Figure 4.2: The sheet music corresponding to the abc-notation in Figure 4.1

### 4.1.2 Irish Folk Dataset

One community that embraced this format is that of Traditional Irish Folk musicians. The website *The Session* [3] is probably the most important there are thousands of different pieces, called tunes, often with numerous variations and comments from the users.

The website provides a `.json` file with the complete database of tunes in abc-notation. This was used by Sturm [122] to create a tokenized dataset which they used to train a deep neural network for music generation, called folkRNN.

[3] `www.thesession.org`

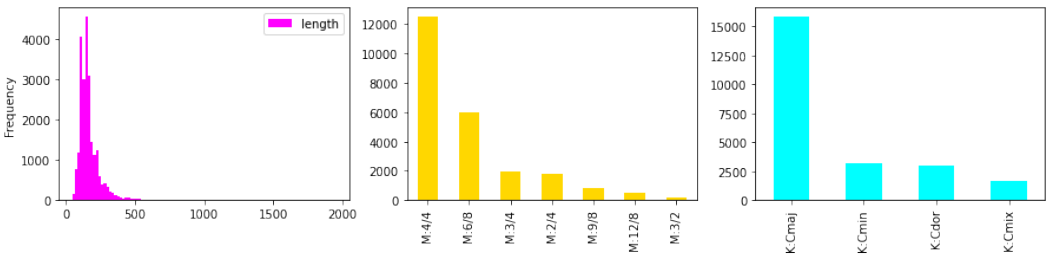



Figure 4.3: Histograms for length, meter and mode. Most of the dataset is Major and 4/4. Average length is 168

The first step of project used the same dataset so that folkRNN could be a point of comparison.

The dataset contains around 30,000 rows, but not all of those are valid tunes. Some of those are only partial "variations" of others already present, others are modern compositions inspired by Irish folk that however contain elements not present in traditional tunes (e.g., odd time signatures, unusual tuplets, chromaticism)

The tokenized version contains around 20.000 tunes with a vocabulary of 137 tokens, including three tokens for start and end of sentence and for padding. All tunes are transposed to C, while maintaining their original mode which for traditional folk is either major, minor, dorian or mixolydian. Time signatures usually correspond to certain type of dances. The most common are Reels in 4/4, Jigs in 6/8, Waltzes in 3/4 and Marches in 2/4.

The distribution of those tunes is however not uniform, with the vast majority of tunes being in 4/4 and in the major mode. The average length is 168 tokens when considering the whole dataset with 2,252 tunes longer than 256 and only 214 longer than 512. The model we used requires the specification of a maximum sequence length which we initially set to 256 as it does not pose problems for the Irish dataset but was subsequently raised to 512 as the Swedish dataset, described in the next section, had longer tunes and we wanted to keep as much of these examples as possible.

### 4.1.3 Swedish Folk Dataset

As we will describe in depth in section 4.4, we also experimented with a different type of folk music in the context of *Ai Music Generation Challenge 2021*. The challenge focused on a particular sub-genre of traditional Swedish dance music called *Slängposlka*.

This music is characterized by a 3/4 time signature and comes either in the major or minor mode. However, minor tunes make frequent use of the harmonic minor mode, that is the 7th degree of the scale is raised in cadential passages giving a peculiar sound. Contrary to the usual rhythmic structure of 3/4 music like waltzes, where the is a strong beat and two weak beats, in slängpolska we find 3 even pulse in each measure which can be divided in two 8th notes or very often into a series of 16th notes.

These characteristics make this genre quite different from the examples we have in the Irish dataset, while sharing the same representation and general concepts behind dance music. Additionally, the available dataset, scraped from the website *FolkWiki*[4], is quite small, only containing around 600 examples.

Our goal was then to create a model for slängposlka that could effectively learn the style from the limited data while leveraging the common knowledge learned from the Irish dataset.

## 4.2 Methods

This section describes the methods used in this project, namely the transformer architecture and the visualization techniques we devised around it.

### 4.2.1 The Transformer Architecture

Transformers are the current state of the art for sequence modeling, they revolutionized natural language processing and are now being used also for computer vision tasks.

This neural network architecture was first introduced by Vaswani et al. [123] in the context of neural machine translation. The intuition is that of making the attention mechanism the main driving force of the model. Previous work already leveraged this architectural feature but, as the title of the paper claims, it may be the only component needed for successful learning. The basic idea is that of learning a series of matrices that enable scoring the relative importance of each token when it comes to computing the output (usually predicting the next token or value in a sequence). To do so we compute three matrices called key, value, and query. They can be thought of as a sort of soft associative memory system that learns what to retrieve, or look at using the attention metaphor, based on the inputs. A common modification to the original formulation of the attention mechanism is that of splitting the latent space to reduce the computational weight,

[4] `www.folkwiki.se`

also allowing to split the attention to multiple aspects of the input, in what's called multi-head attention.

The equations below describe the actual computation, where $Q$ is the query, $K$ is the key, and $V$ is the value; $d_k$ is the dimension of the embedding.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4.1}$$

Another important part in the transformer is the positional encoding. The attention mechanism in fact, is not able to differentiate between the same token in different positions as they would share the same embedding vector. RNN do not face this problem as the inputs are presented one after the other, making their positional context explicit in the way the go through the network. Transformers, on the other hand, process everything at same time, so to explicit the positional information and encoding function is used, which is summed to the embedding vectors according to the following formula. Successive timestep will sum different values to each dimension in the embedding space.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{n^{\frac{2i}{d_{model}}}}\right) \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{n^{\frac{2i}{d_{model}}}}\right) \tag{4.2}$$

Apart from these two peculiarities, transformers only feature feedforward layers with residual connection in the attention block. Once the inputs are encoded, they go through a number of blocks and the final representation is used to compute output probabilities. On a macroscopic level we could then say that RNN and transformer are equivalent with slightly differing "memory" mechanisms [124].

### 4.2.2 A Model for Traditional Music: The *Tradformer*

Starting from an open-source implementation[5] of OpenAI's GPT-2 [125] we developed a transformer model specifically optimized for symbolic music generation that we called Tradformer. Figure 4.4 shows the architecture of the model with the different parts each colored in a certain way.

The hyper-parameters where chosen through human-in-loop evaluation of the outputs. Same goes for the choice of the positional encoding, which was a learned vector in the original implementation which we reverted back to sinusoidal function proposed by Vaswani. Sampling proved to be the most challenging part that required a lot of effort and human expertise; the algorithm we devised is documented in Section 4.3.

[5] https://github.com/karpathy/minGPT

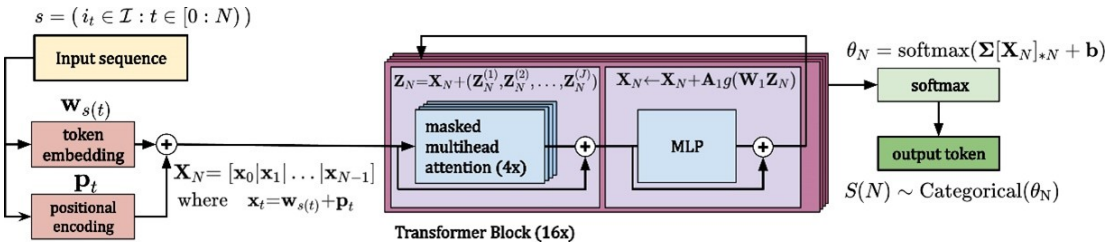



Figure 4.4: Tradformer architecture

### 4.2.3 Visualization

After we realized the fact that the transformer was under-performing out of the box, we started an iterative process of development and refinement which is described in Section 4.3. Visualization played a fundamental role in guiding our intuition and bringing together the musical knowledge and the machine learning knowledge.

#### Softmax

Let us start by considering the softmax visualization. As the name implies, we plot the softmax scores that the model outputs at each timestep. These scores are very useful because they show what the model has learned as a probable output. Also, when using a sampling strategy that draws from a distribution it allows us to see if mistakes are due to a low probability token or if the whole distribution is skewed towards values that are not correct.

```
X:0
M:6/8
K:Emin
|:G|EAA ABd|ege edB|AGE EDE|G3 GFG|EAA ABd|ege edB|AGE EDE|GAG A2:|
|:d|eaa bag|ege edB|AGE EDE|G3 GFG|eaa bag|ege edB|AGE EDE|GAG A2:|
```

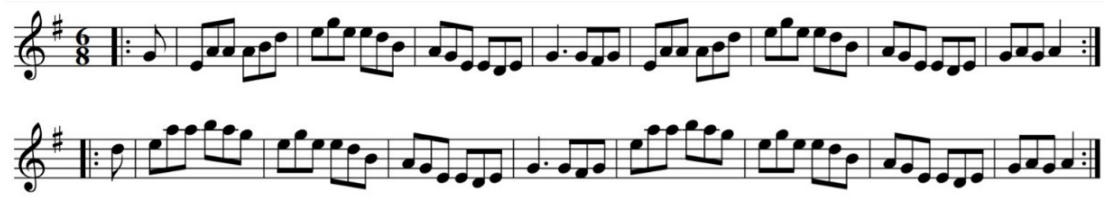


Figure 4.5: An example output from the Tradformer, transposed to E minor. This tune is the same that appears in the softmax plot in the next figure

Figure 4.5 shows the abc and musical notation for a reel generated by the Tradformer. This reel is not particularly interesting, but it exhibits the regular structure often found in this type of tunes.

The A-part start with statement of a theme in the first four bars that is reprised almost verbatim in the next four bars with the exception of the last one. This type of structure is very common. The only strange thing about is the fact that the tune ends on an A, which is not a note present in the tonic chord of E minor (which contains E, G and B). The B-part starts with a different pattern in the first bar but still follows the A-part in the subsequent three. The second half of the section is identical, with the last bar choosing the same unusual note as before. When we take a look at the softmax plot in Figure 4.6 we find an explanation of this behavior.

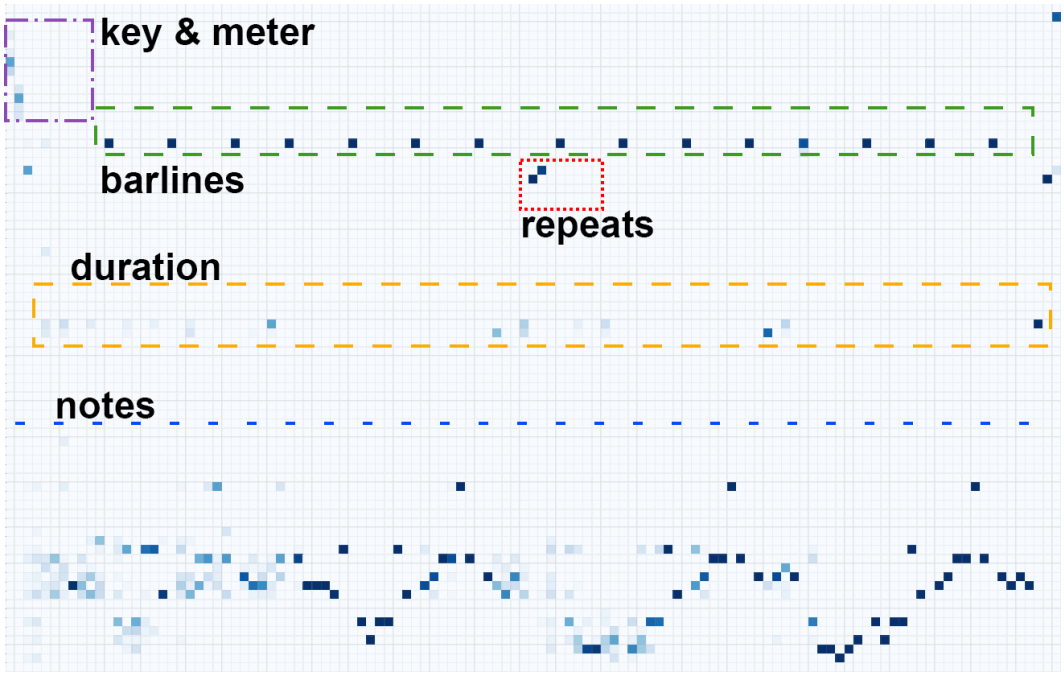



Figure 4.6: Softmax visualization. Columns correspond to a timesteps and rows to tokens in the vocabulary. Darker values correspond to higher probabilities

The model starts off with a probability distribution that is quite uniform between the note tokens. The same uniformity can be seen at the start of the B-part after the repeat sign. This means the model has learned that at the beginning of each section the melody is not biased towards anything in particular.

After a couple of measures however the probabilities appear to be highly concentrated on specific tokens. This is because the model has learned that in this style, thematic ideas in the melodies get repeated after two or four measures and so it looks back at the previous steps and chooses the same tokens. Between the A and the B-part we can see that the model is a bit uncertain on how to end the section. That unusual note we saw in the finished tune actually has a

lower probability than the G token, which would finish the section on the 3rd of the tonic chord a sound more familiar and perhaps satisfying. It is interesting enough to notice that the model has learned that the ending of the A-part is often reprised for the B-part and so when it comes to the last tokens it assigns almost 100% probability to the same sequence we have seen before even if they had low probabilities before.

### Weights of the Final Layer

Plotting the weights matrix of the last layer of the model, often called the classification head, is interesting not only as a debug tool, like the previous one, but also as it provides confirmation that certain musical concepts were learned.

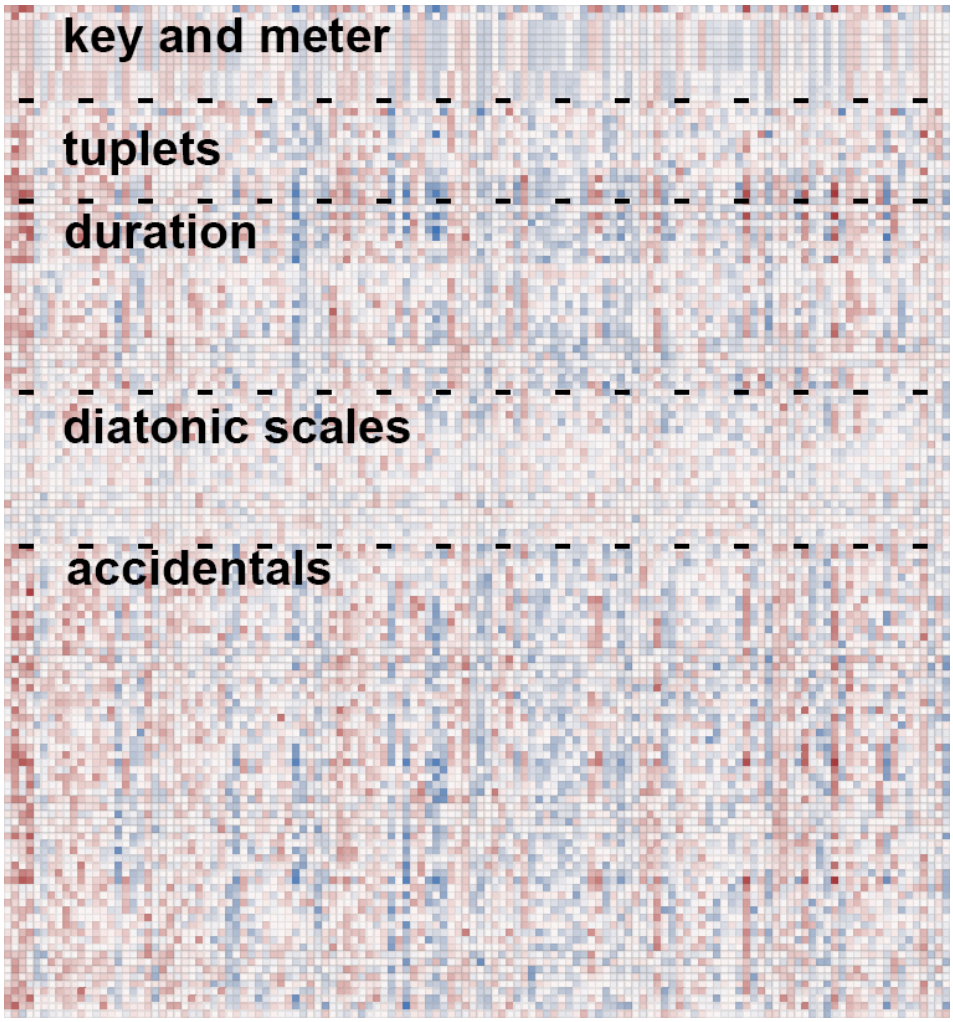



Figure 4.7: The weights of the last layer of the model.

This layer is responsible for translating the latent information at the end of the model into a probability distribution for the outputs through the softmax function. In Figure 4.7 each row corresponds to one of the 137 output tokens and each column corresponds to the 128 dimensions of the latent space. Red hues represent positive weights and blue hues represent negative weights, with white around 0. However, the actual weights are not interesting to us, but rather the evident similarities we can see when we sort the rows by the musical function of each token.

The first group we can see at the top is the "additional" token for start and end of sequence and for padding; after that we can clearly see two bands: one for the meter tokens and one for the mode tokens.

Below that we have a group of "structural" tokens, like measure bars and repetition marks, and then all the token that relate to tuplets. Here we noticed a strange phenomenon: tuplets beyond triplets (which are quite common in some style of Irish music) share a very similar weight vector to certain duration tokens which are the block right next to it.

Our hypothesis is that those tokens have in common the fact of being rarely used, which we verified by looking for them in the dataset, and their weights are thus probably similar for this reason. When we trained the slängpolska model we removed those without any loss of performance, and we could argue that their presence was not very sensible given the styles we are considering.

### Embedding Self-Similarity

Closely related to what we just observed is also the plot of the self-similarity matrix of the token embedding. This plot gives us a nice visual confirmation of the fact that the model has learned certain musical concepts and relationships that we could already suspect from the previous plot. It is also interesting to notice that those same structures also appeared in the weight matrices for the gates in folkRNN [126].

In Figure 4.8, each row and column correspond to one of the 137 tokens. Hotter colors towards red represent high similarity (1.0), colder colors towards blue/green indicate opposite vector (-1.0) and yellow indicates independent vectors (0.0). Like before, we can see clusters forming that correspond to certain musical functions (keys, meters, duration, etc.). We also see a nice pattern in the similarity between notes of diatonic scale (notes without alterations, the white keys on the piano) across different octaves, indicated by diagonal lines in the plot. Similarity also between the natural and altered version of each note also appears, with blocks that seem to relate to flats, sharps, and naturals at the bottom-right end of the

plot.

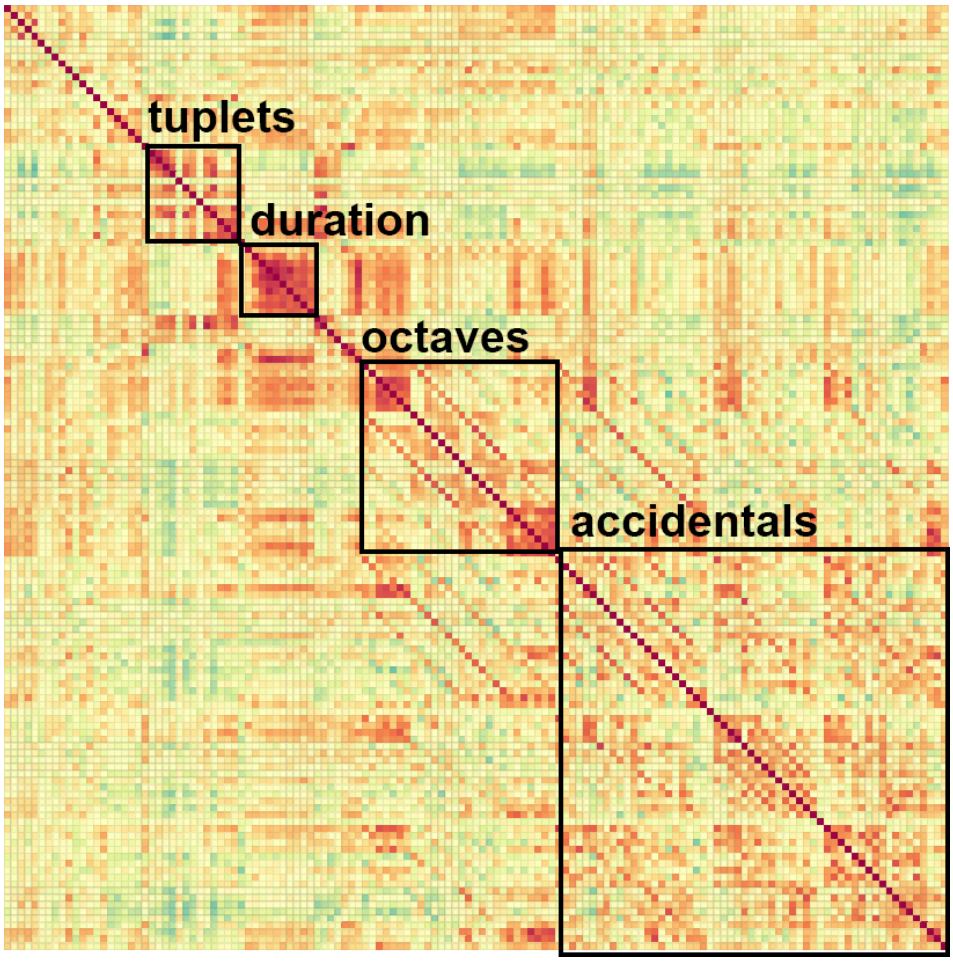



Figure 4.8: Self-similarity matrix of the embedding layer. We can see clusters and lines corresponding to specific musical relations between tokens.

### Attention Heads Scores

Finally, we visualized the attention scores for each of the 4 heads in each of the 16 layers. Plotting the product of the key and query matrices gives us the possibility of visualizing what each attention head is looking at.

Figure 4.9 shows the attention scores at layer 1, 8 and 16 for the same tune used in the Softmax plot and shown in figure 4.5. Here darker values correspond to higher attention scores for the token at that timestep in the input sequence; those scores and used to compute the output at the next timestep. This kind of plot has become very common in transformer applications and is useful to get an idea of

the important part of the input. However, some researchers have pointed out how this visualization is not a good form of explanation [127]. This is because after the first attention block, the fully connected combine and shuffle all input position together, making the plots of subsequent layer unrelated to the input sequence. Nonetheless, certain structural pattern can be seen in the attention plots that are definitely connected to regularities in the inputs, even if their location may not be completely informative.

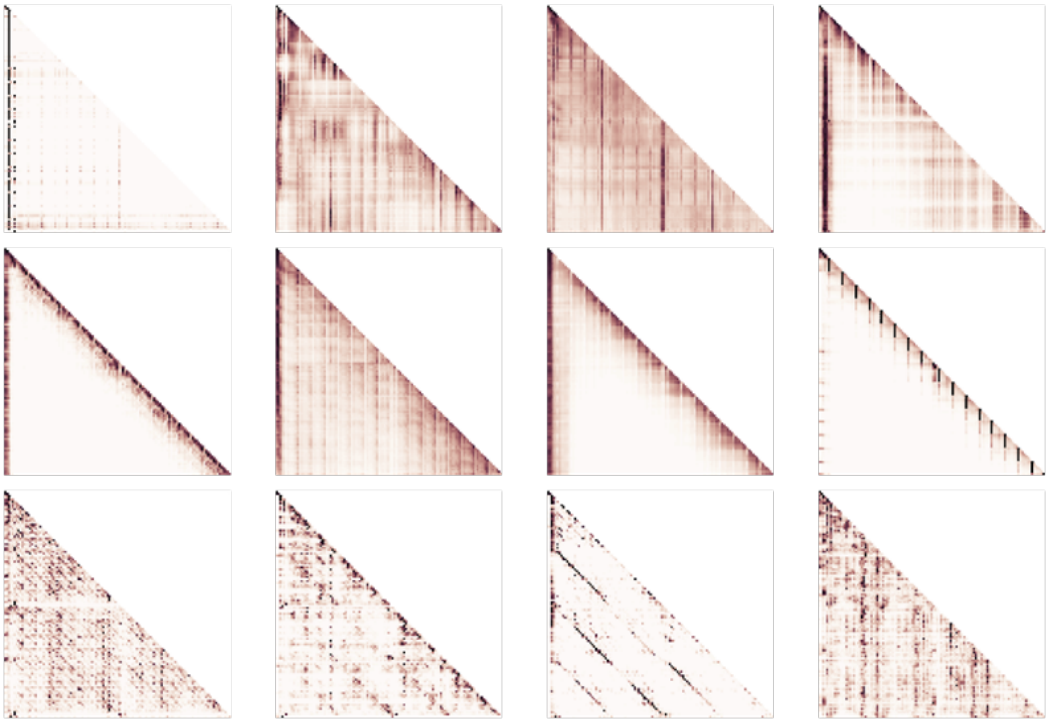


Figure 4.9: Attention plots for the first, eight and last layer on the Tradformer. We can see patterns that are reminiscent of musical structure

In the first layer we can see (this also apply to the next few) we see attention patterns that are either very focused on the initial tokens or are very diffuse. This makes sense as the initial tokens are the one that contain information about selecting the right notes and rhythms (key and meter tokens) which are the most basic concepts and are very important for generating a correct piece.

Towards the middle layers of the model, we start to repeated structures at regular intervals. While, as stated before those are not necessarily related to the position they are in, they unmistakably relate to measures as they match their number. We interpret this as an indication that the model is now focusing on higher level concepts like the melodic content of each bar.

Once we reach the end of the model, we see very sparse patterns which could related to any number of things, but it is interesting to notice how we find diagonal structures which resemble the bipartite structure of the tune, with an A and a B section each composed of two phrases in a call and response scheme.

### 4.2.4 Sampling Strategies

The sampling strategy of Tradformer is extremely important. We found the biggest improvement in the quality of the generated music came from replacing a naive approach with a more sophisticated one based on beam search.

Our early models were using a naive sampling approach, where single tokens were drawn from a categorical distribution parametrized with the softmax probabilities. The outputs contained counting errors and drifting melodies and were rated very poor against melodies generated by folkRNN.

Lowering the softmax temperature was slightly helpful in preventing such problems but created too much repetition. Furthermore, temperature is applied equally even when the context makes it unnecessary.

We thus tried employing top-k and top-p sampling [128], where unlikely tokens, according to rank or probability mass, are removed before the remaining probabilities are re-scaled by softmax. With $p \in [0.9, 0.99]$ we saw fewer mistakes but also less variety. This could be countered by increasing the temperature above 1.0, but we observed the melodies were still drifting. Even a nudge to these hyper-parameters led to wildly different results and finding the right balance was difficult.

This led to our development of our own flavor of beam-search sampling. We employed a combination of beam search and nucleus sampling similar to that proposed in [129]. Tradformer starts with top-p sampling (only looks at the most likely token up to p percent of the probability mass) and then uses at most k of those as branches to search a tree of depth of maximum depth $D$. This results in a sample space of at most $k^D$ token sequences with a probability distribution given by the softmax of the sum of the logits of the component tokens. This increases computational cost, but the increase in output quality was clear. Tradformer sets $D = 3$, $p = 0.99$ and $k = 3$.

Beam search provided melodies with more direction and virtually no mistakes but, as noticed for NLP in [128], makes for less "surprising" outputs. Correctness and variety seem to be two different objective that are difficult to obtain at the same time. However, leaving some control over those parameters to the user can provide creative opportunities.

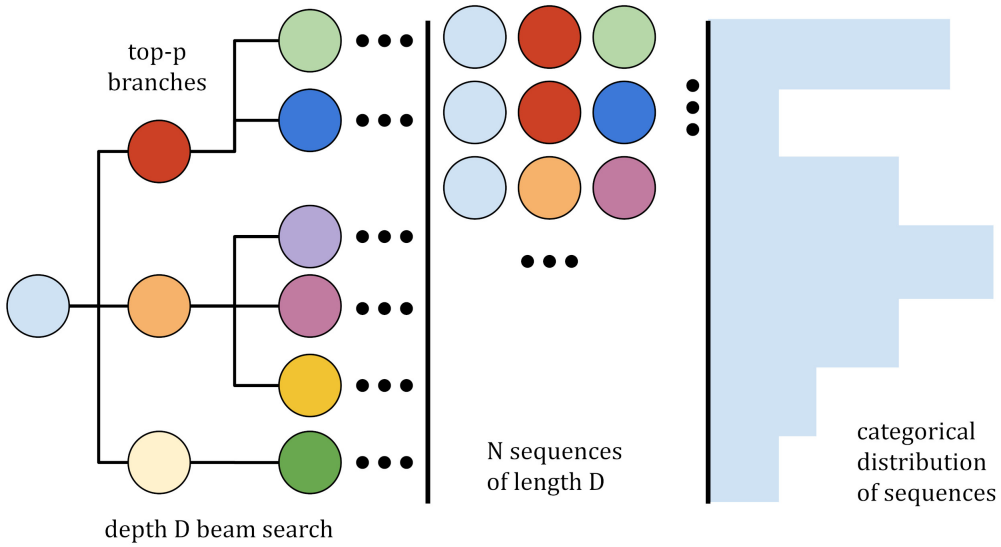



Figure 4.10: Schematic representation of the proposed sampling strategy

## 4.3 Matching the Performance of folkRNN

The open-source implementation of the transformer model could work properly with the folkRNN dataset as it came, by just specifying the correct vocabulary size and using the tokenized tunes. We expected this state-of-the-art architecture to perform well enough right out of the box, but this was not what happened.

The model was making considerably more mistakes than folkRNN ever did, and this unexpected issue led to several iteration of the model which eventually become the Tradformer. This iterative development was guided by human expertise, both as machine learning experts and musicians.

We setup a series of human-in-the-loop evaluation experiments using the expertise of the data scientists in the MUSAiC project who are also musicians. The experiments were conducted as follows: 10 tunes would be generated using the same prompt by both the tradformer and folkRNN. Any information that could help recognizing the output like the way tokens are grouped would be hidden and the tunes would be randomly shuffled. After that an evaluator would look at them and score them with a simple system (good or bad) while also providing a brief comment.

Figure 4.11 shows the first three rounds of evaluation. As we anticipated the transformer model was performing badly out of the box in round 1. Adjusting the hyperparameters using our expertise and the insight from visualization, find out

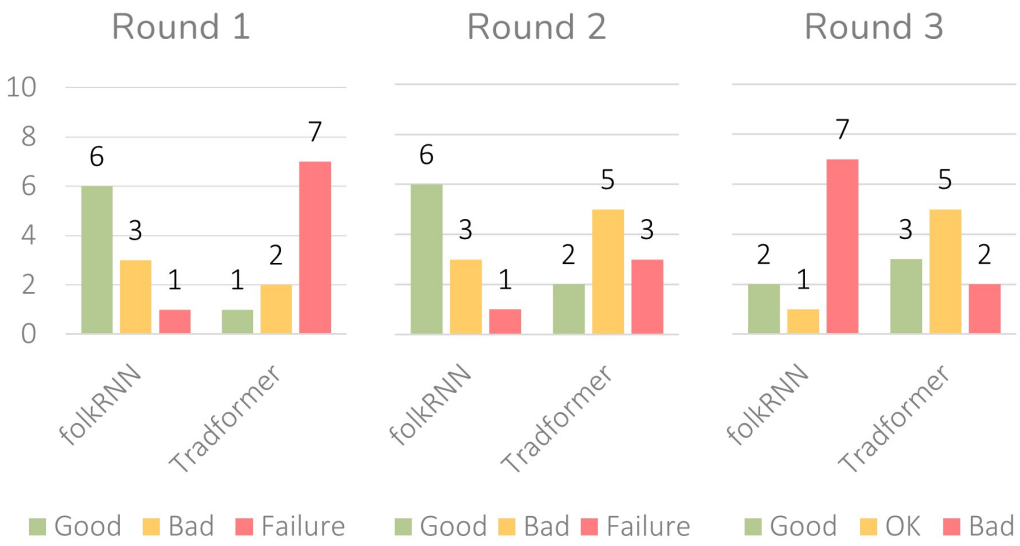



Figure 4.11: Human-in-the-loop evaluation

that we did not need as many attention heads and that the latent space needed not to be very big, yielded a better model for round 2 but it still performed poorly in comparison to folkRNN. Round 3 saw the introduction of a better sampling strategy once we realized how susceptible the model outputs are to that. We continued with further evaluation focusing on specific styles (4/4 jigs and 6/8 reels) where the tradformer still came out on top so we considered our development process successful.

## 4.4 Transfer Learning on Slängpolska: The Ai Music Generation Challenge 2021

After matching folkRNN performance we wanted to test the flexibility of this model by employing it in a generation task with a small dataset and a slightly different musical language. The *Ai Music Generation Challenge 2021* provided the perfect chance to test our model capabilities against other approaches to the same task and also to see how human experts would evaluate its outputs.

The challenge consisted in generating 1000 tunes in a particular style of Swedish dance music called Slängpolska. There were no rules concerning the type of models or computational resources to be used. The organizers pointed the participants to a dataset of original tunes which we described in section 4.1.3.

After the submission 9 tunes would be selected at random form the 1,000, and an additional one would be handpicked by the participants. After that, a number of judges, all of whom were either expert musicians or dancers of traditional Swedish music, would score the tunes according to four parameters.

**Formal Coherence**: the adherence proper musical formalism, such as the correct use of repetitions or the absence of unwarranted missing/additional notes in each measure.

**Stylistic Coherence**: the adherence to the mannerism of Slängpolska and how well this was captured by the model

**Playability**: how well the generated tune can be performed on a traditional instrument. These usually have limitation when it comes to range and articulations.

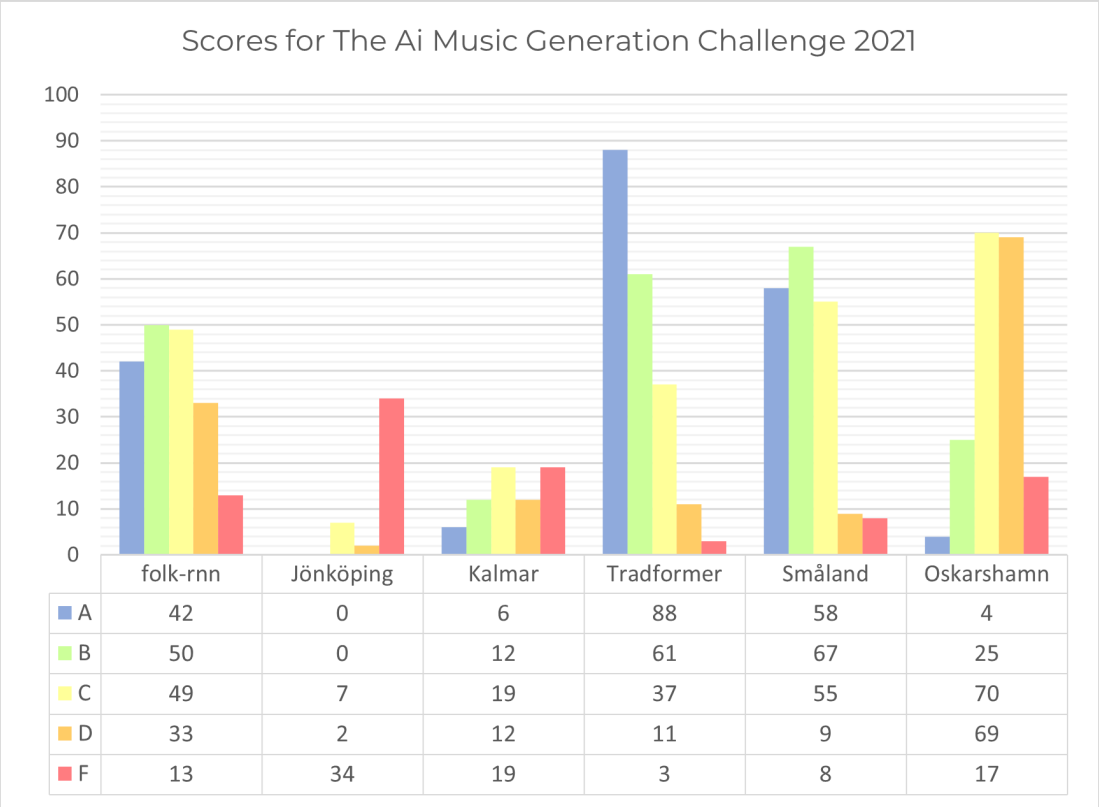
**Danceability**: how well the tune could be performed by dancers, needing a regular rhythmic structure in each measure and in the tune overall. Tunes exhibiting plagiarism or presenting meter or rhythm uncharacteristic of Slängpolska would be discarded and receive a null score (F)

The Tradformer was adapted as follows for the task. Firstly, we increased the maximum sequence size of to 512 tokens as Slängpolska in the dataset tended to be longer than Irish music and we did not want to lose any available information. We also removed all the rare and unnecessary tokens from the vocabulary following the intuition we got from the weight matrix plot in section 4.2.3. This lowered the vocabulary size from 137 to 128.

We observed that the outputs of the tuned model suffered from problems that can be traced back to the lack of data. Most notably, certain patterns tended to be repeated too often. The model had a tendency to get "stuck" in cycles of two bars – a behavior not seen in either the Irish or Swedish datasets. We also found miscounted measures. The Swedish tuned Tradformer would create odd-length sections, but not as often as seen in the Swedish dataset. Furthermore, we found it hard to balance between interesting output and the risk of too much repetition.

Hallstrom et al. [130] report that fine-tuning a folkRNN model on the whole FolkWiki database (not just Slängpolska) also presented difficulties in adaptation, and generated melodies described as unfocused. The strict policy on malformed tunes motivated us to employ rejection sampling to filter out generated material that would score nothing.

We devised a series of conditions and generated new tunes until the model produced the thousand tunes required in the challenge submission. Inspired from the analysis by Sturm and Ben-Tal [131], we formed criteria by looking at statistics




| | folk-rnn | Jönköping | Kalmar | Tradformer | Småland | Oskarshamn |
|---|---|---|---|---|---|---|
| A | 42 | 0 | 6 | 88 | 58 | 4 |
| B | 50 | 0 | 12 | 61 | 67 | 25 |
| C | 49 | 7 | 19 | 37 | 55 | 70 |
| D | 33 | 2 | 12 | 11 | 9 | 69 |
| F | 13 | 34 | 19 | 3 | 8 | 17 |

Figure 4.12: Final scores for the AI music generation challenge 2021

from the original dataset and choosing reasonable values. One condition was that all bars must have the correct number of notes. Another condition was that a tune does not contain too many repeated measures, and in particular too many repeated couples – as this type of repetition happens frequently in the generated material but not in the training data. We also checked the tune length and pitch range, along with the number of bars in each section. Very short or lengthy sections show a generation that has gone amok. Rejection sampling discarded around one tenth of the outputs, evidencing the baseline quality of the model.

The Tradformer ended up winning the contest, surpassing the performance of folkRNN and other LSTM based models used by the other participants. Figure 4.12 shows the overall scores.

The following Figures 4.13 to 4.22, show the nine tunes that were randomly selected for the challenge plus the cherry-picked one, in traditional music notation.

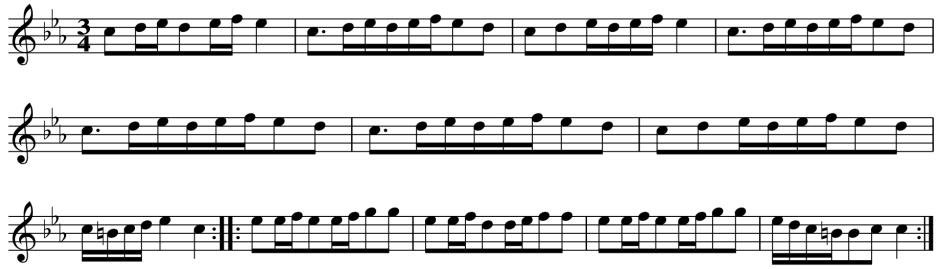


Figure 4.13: Tune #108

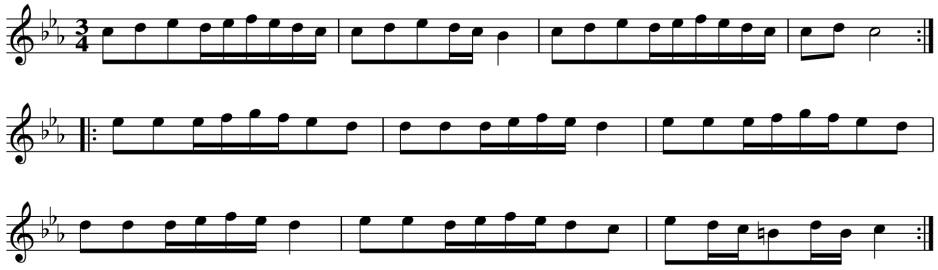


Figure 4.14: Tune #117

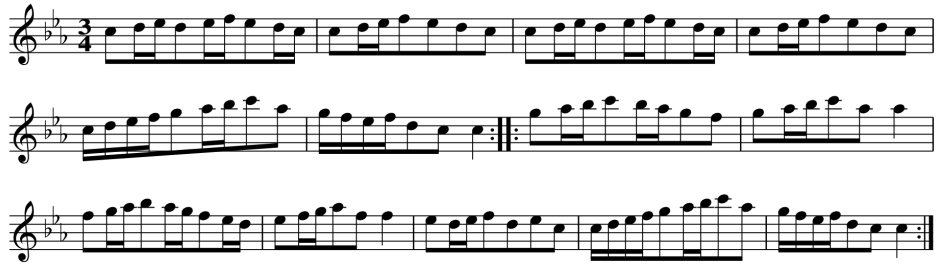


Figure 4.15: Tune #263

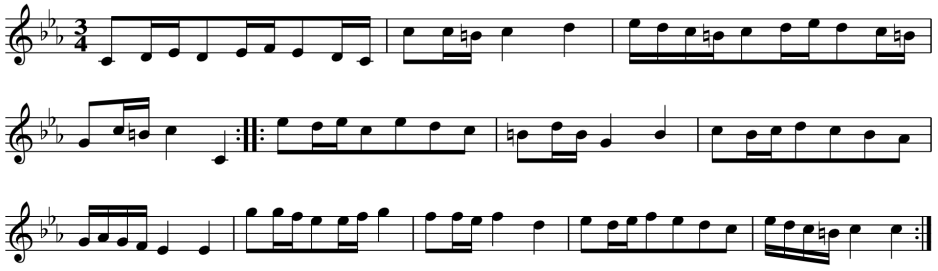


Figure 4.16: Tune #267

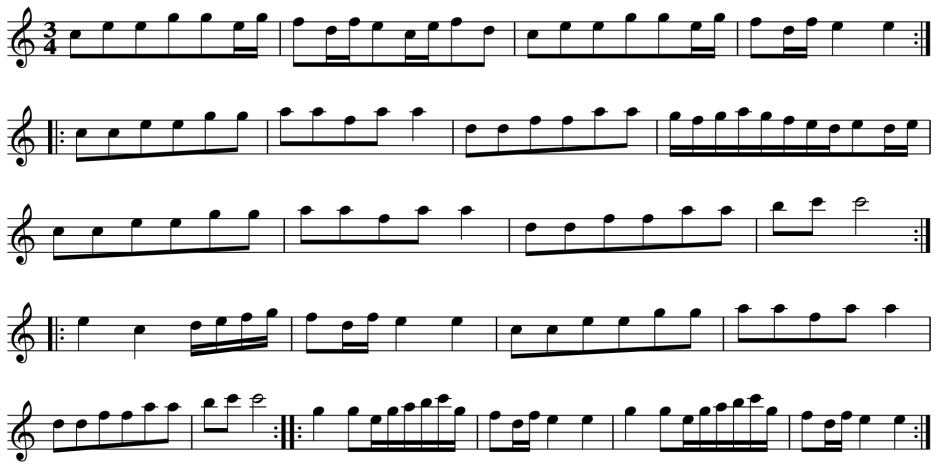


Figure 4.17: Tune #463

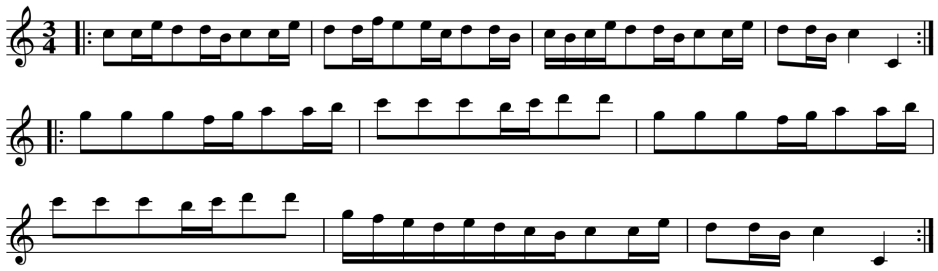


Figure 4.18: Tune #553

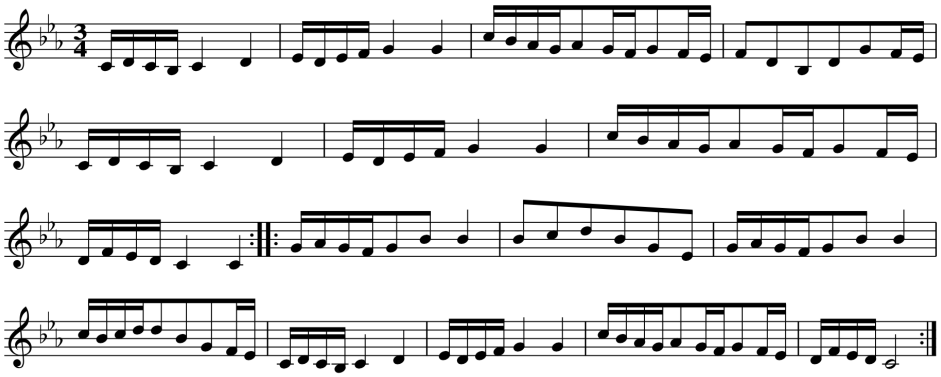


Figure 4.19: Tune #576

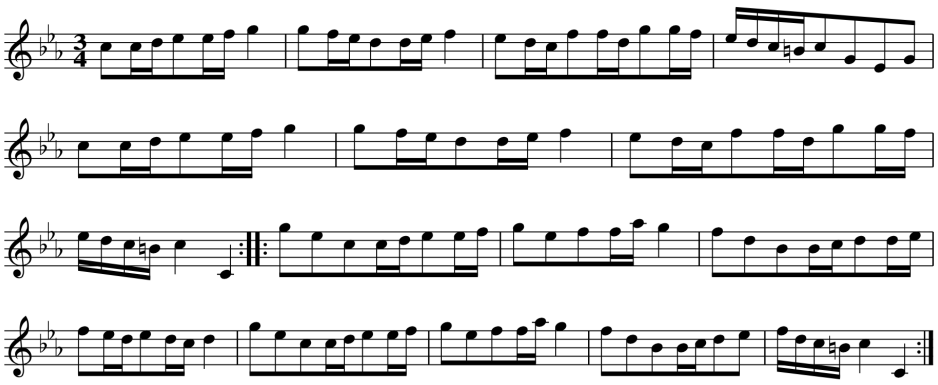


Figure 4.20: Tune #738 (hand-picked)

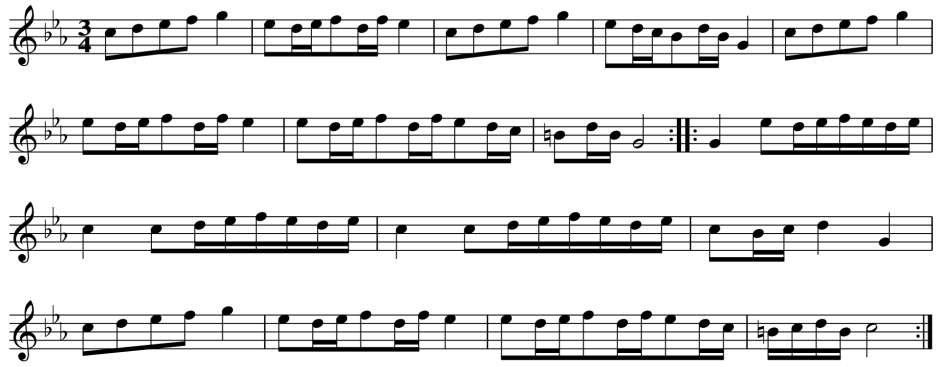


Figure 4.21: Tune #751

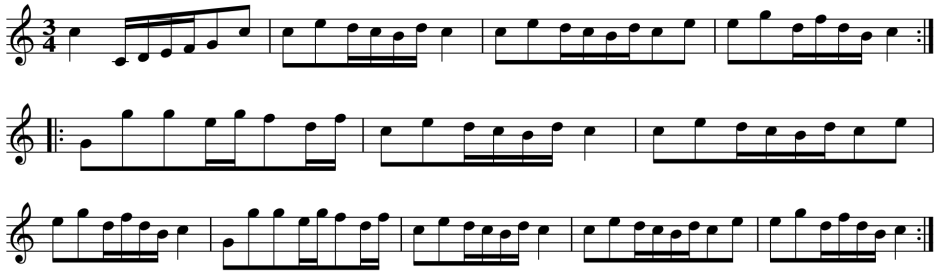


Figure 4.22: Tune #900

### 4.4.1 Scores and Comments from the Judges

The following 5 tables contain the scores and comments for each tune from each judge. It is interesting to see how different judges were more or less lenient towards certain quirks of the model like the odd structure it produced sometimes. It could be an indication of a kind of bias towards AI technology in the sense that the expectation is lower if we know that a machine is producing such artifacts. On the other hand, the fact that these quirks are found "interesting" by some may reflect their creative inclination and acceptance of things that push the envelope of traditional forms, appreciating the potential of such a model to provide examples of a familiar music with a different spin.

Nonetheless for the best scoring tunes we see a recurrent comment praising the authenticity of the composition, signaling how the model has effectively grasped what makes a tune traditional. The irregularities and "boringness" are two faces of the same medal and go back to the problem of balancing the errors and repetition in the sampling phase.

Table 4.1: Judge 1 scores and comments

| Tune | D | S | F | P | Comment |
|---|---|---|---|---|---|
| 108 | B | C | C | B | Very repetitive in the A-part |
| 117 | A | A | A | A | Typical Slängpolska in minor. Quite simple but in the style. A motif comes back in the end of B-part. |
| 263 | A | A | B | A | A bit odd to repeat the first two bars. Nice with the different length of the phrases in the B-part. |
| 267 | A | A | A | A | Going back to the Bb in measure 7 seems a bit sudden, but it still works in the context I think |
| 463 | B | B | C | B | There are some good elements melodically and rhythmically. The length of the C part is 6 bars makes the tune a bit unbalanced |
| 553 | B | B | B | B | It's a bit confusing in the beginning because you feel it as 4/4 but then it "resolves" to $\frac{3}{4}$ again. I like this melody. It makes it interesting when playing with the meter |
| 576 | A | B | A | A | Not a typical Slängpolska melodically but a good melody and rhythm. Theme of the A-part comes in again at the end of B-part. |
| 738 | A | A | A | A | |
| 751 | A | B | B | A | Not a very common key. A part end surprisingly on G. |
| 900 | B | B | B | A | A bit repetitive, but the rhythms are good. |

Table 4.2: Judge 2 scores and comments

| Tune | D | S | F | P | Comment |
|---|---|---|---|---|---|
| 108 | C | C | C | C | It can be difficult to feel the first beat in the bars with all repetitions of the little phrase. Just one motif is repeated with small variations in the first part which sometimes can be suggestive or maybe a bit boring. |
| 117 | B | B | B | B | The tune reminds of an older type of Slängpolska. Feels familiar |
| 263 | B | B | B | B | It's not so common with 7 bars in a repeat. The dancers often want even bars. The repeated motif in the 1:st part can give a suggestive feeling. Bar 6 in 2nd repeat doesn't fit so good. I miss one bar in the end of part 2. Cm can be sometimes tricky on the violin. I tried it in Am. |
| 267 | A | A | A | A | Easy to learn and remember. "Typical" tonality. You get the feeling that you can find this tune in a spelmansbok. |
| 463 | D | D | D | D | The 3rd part is not coherent. The two first parts works |
| 553 | A | B | A | A | I played in G |
| 576 | A | B | A | A | Gives nice energy for the dance. Nice tune. |
| 738 | A | A | A | A | This tune feels genuine. |
| 751 | A | A | A | A | The repeat of the motif in bar 6 and 7 in the second part is nice. Nice polska. The tune feels authentic. |
| 900 | B | B | B | B | |

Table 4.3: Judge 3 scores and comment

| Tune | D | S | F | P | Comment |
|---|---|---|---|---|---|
| 108 | C | C | D | C | repetitive but some nice variations of the simple motive |
| 117 | C | B | A | C | simple but effective, coherent |
| 263 | C | B | A | C | Quite good. Nice sequences |
| 267 | B | A | A | B | |
| 463 | C | C | D | C | OK, strange repetitions |
| 553 | C | A | A | B | simple but effective, coherent, |
| 576 | B | B | A | B | nice simple melody, very convincing |
| 738 | B | A | A | B | |
| 751 | B | A | B | B | Simple but coherent. Very repetitive but in a nice coherent way |
| 900 | C | B | D | C | very repetitive |

Table 4.4: Judge 4 scores and comments

| Tune | D | S | F | P | Comment |
|---|---|---|---|---|---|
| 108 | C | C | B | B | Too many repeated phrases, small variations though. |
| 117 | A | B | A | A | Feels like a tune, but a little boring one. |
| 263 | C | C | D | A | Feels better to play in G or D minor for example. Nice tune, if wasn't for the uneven measures, 6+7, funny but not in style, at least unusual. |
| 267 | C | C | B | C | Rhythmically good, ok melody. |
| 463 | F | D | D | C | Some phrases could fit but too many repetitive phrases. Four parts is unusual |
| 553 | C | D | B | C | Sounds like a Danish tune or a menuette. Cool. |
| 576 | A | B | A | A | Nice tune. E-minor felt good to play in. |
| 738 | C | C | A | A | More *1-3* feel in the rhythm but still even. Feels like a tune, maybe not the most brilliant one but still. Better played in d minor. |
| 751 | C | C | A | A | More *1-3* feel in the rhythm but still even. Feels like a tune, maybe not the most brilliant one but still. Better played in d minor. |
| 900 | A | B | A | A | It's a nagging tune. |

Table 4.5: Judge 5 scores and comments

| Tune | D | S | F | P | Comment |
|---|---|---|---|---|---|
| 108 | C | F | F | B | lacks melodic coherence and development |
| 117 | A | A | A | B | limited material, but works |
| 263 | B | C | B | B | a bit unusual, but also a bit interesting |
| 267 | B | A | A | A | polonäs style OK |
| 463 | A | A | A | B | |
| 553 | A | A | A | A | could be traditional! |
| 576 | A | A | A | A | nice, in the box, nice dorian in the B part! |
| 738 | A | A | A | B | makes tonal sense - even a progression in the second part |
| 751 | A | A | A | A | good! within the box, makes tonal sense - even a progression in the second part |
| 900 | B | A | B | A | lacks formal contrast, but within the box |

## 4.5 Music Co-Creation

In this section we want to describe how the model has been used for music co-creation sessions, where the outputs served as the basis for further composition. In this scenario the model becomes an assistant, an extension to human creativity [132–134].

Traditional folk music already provides for the possibility of performers including their own embellishment and variation to a tune, as long as the melody is still recognizable. For this reason, the way the dataset was tokenized completely ignores any symbol that is related to expression and only conserves the actual notes. When a practitioner plays the outputs of folkRNN or the Tradformer, they will "fill in the gaps" by introducing their sensibility to the playing anyway. In this sense, the lack of detail of the outputs can be seen as a feature rather than a shortcoming.

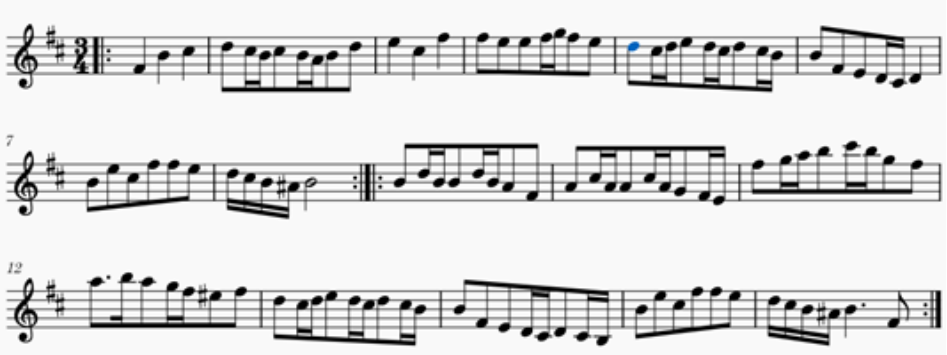


Figure 4.23: The original output from the Swedish Tradformer

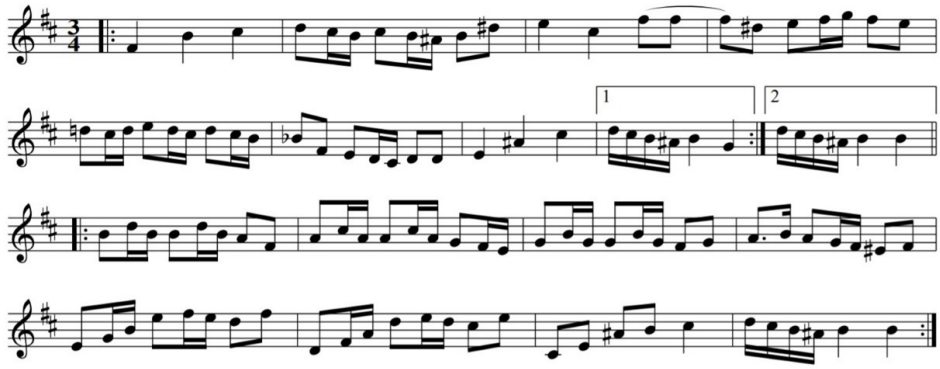


Figure 4.24: The revised version of the tune we title Ugglas Polska. The A-part sees the introduction of tonal ambiguity and the B-part is partly rewritten to extend a passage we were fond of.

Figure 4.23 shows one of the outputs of the Swedish model and Figure 4.24 shows the final version of the same tune, with multiple tweaks we introduced while playing it we describe below:

- Transposition from C minor to B minor, a more appropriate key for the instruments we played.
- Introduction of some tonal ambiguity through alternate use of the notes D and D# in bars 2,4 and 5.
- Slightly alteration of bar 6 to include the leading tone A#.
- The phrase in bar 7 was replaced with an arpeggio outlining the dominant chord resolving in bar 8.
- Bars 11, 13 and 14 were rewritten to develop and vary the pattern presented in bars 9 and 10.
- Bar 12 transposed down an octave in order to avoid jumps and continue the phrasing.
- Bars 15 and 16 rewritten to revisit the ending of the first part to create a satisfying conclusion.

Such adjustments to generated material, which are not always necessary fixes but rather aesthetic decisions, can be part of the workflow of a composer. More examples along with recordings can be found online [6].

## 4.6 Conclusion

This chapter documented the work done as a visiting PhD student taking part in the MUSAiC project. The set goal was exploring the differences and similarities between transformers and recurrent neural networks when applied to the task of symbolic music generation. In particular, the work focused on learning how to generate traditional folk music from Irish and Swedish traditions and to compare the results with the previous state of the art, folkRNN.

At the beginning, our expectations were that the transformer would obtain extremely good performance without much effort, given how since their introduction in the world of natural language processing they have redefined state-of-the-art performance in every of the major tasks. However, we encountered many obstacles along the way and the quality of the outputs was consistently worse than those generated by folkRNN.

[6] https://tunesfromtheaifrontiers.wordpress.com/

To address the problem, we started tuning hyperparameters in order to lower the model loss, but we soon found out how the metric was not telling a lot about what humans care about: the quality of the generated music. This is understandable, as the loss function does not integrate any concept of musicality and a wrong token always has the same weight, even if sometimes swapping a note for another has no effect on the perceived musical quality. Once again, we have an answer to our **Research Question 3**, clearly showing how metrics are something to treat carefully and that does not always align by default with the human concept of value.

To improve our model, we thus resorted to human-in-the-loop evaluation, iteratively improving on the model after reviewing the scores given to the outputs by a domain expert (which in this case was ourselves as all MUSAiC research are also musicians). Another key element in this development process was visualization. Looking at how the model generated its predictions and at the inner state of the model gave us precious insight on what was working and what was not. Interpreting this information both as musicians and deep learning experts ultimately led to the creation of an improved sampling strategy. This address both **Research Question 1 and 4**.

In the final section we showed a few examples of human-AI collaboration in the process of creating new music. We started from the outputs of the Tradformer and collectively decided on adjustments that would fit our taste. Sometimes those adjustments addressed an aesthetic want of the musicians while some other times they were necessary to correct mistakes in the outputs that blemished and otherwise enjoyable tune. In the context of **Research Question 2** this shows how in creative application even a less-than-perfect system can kick-start creativity and be of good use to artists.

# Chapter 5

# Observational Studies During the Pandemic

This chapter will focus on observational studies we performed during the pandemic to try to contribute to the scientific discussion and highlight certain issues [135–141]. While the contents of the following section may be seen as a slight departure from what was discussed up until now, there a few considerations that tie together this chapter with the data-centric and human-centric themes previously covered.

First of all, as we will discuss in the next section, the data collection during the pandemic has suffered from the sudden necessity of standardized and capillary infrastructure for diagnostic testing that resulted in datasets that were adjusted overtime with correction to both the methodology and the data itself. Additionally, every country opted for slightly different strategies to deal with the pandemic and for collecting data on the spread. For this reason, we opted for traditional statistical methods that are very clearly interpretable in what they are showing and make no effort of hiding this uncertainty and variance in the various dataset we used. We could say that our methods are data-centric in the sense that their choice is dictated by the only data available during the pandemic.

Traditional methods, with their interpretability, are also important to the human-centric discussion. The pandemic has affected, and continues to affect, the lives of millions around the world. In the midst of this crisis many ideas and hypothesis circulated, some more cautious and other more bold. We felt that observational studies with clear goals and interpretable methods would be the most useful to the public discussion, even though they demonstrated very little. In any case, the lack of data and the difficulty of intervening and experimenting with policies made most of the work on the pandemic, that claimed any degree of certainty, speculation at best. So, we can say that in a context like this, with

human lives at stake, using clear and interpretable models and being transparent about their limitations is human-centric as it considers who the information is going to be consumed by, and how it could affect them, both on an individual and collective level [142].

## 5.1 Data Sources

Here we will describe the source of the various dataset used in the analysis detailed in the next sections and we will also provide some comments on their nature and their usefulness.

### 5.1.1 Covid Timeseries Data

We used time series data for COVID-19 daily new cases and daily new deaths for each Italian region in section 5.2 and 5.3 while we used nation-wide data for a number of different countries in 5.4 and 5.5.

#### Italian Regions

Data for the Italian studies comes from the official repository managed by the Italian department of civil protection (PCM-DPC)[1]. This repository is available on GitHub and is updated daily with the numbers coming from each regional health department independently. This led to discrepancies that were addressed to the best of their capabilities, but not always fixed. Looking at certain regions, especially the smallest one, we can sometimes see negative numbers of cases that signal some kind of recounting has happened due to mistakes.

The procedures behind the collection of some data also changed overtime, most notably for what counts as a positivity test. All these considerations make the dataset to be taken with a grain of salt. However, it must be noted how at the beginning of the pandemic things were moving quickly and this data was all that was possible to obtain.

#### Countries Around the World

The international data was taken from the repository maintained by Johns Hopkins University that collected and aggregated information coming from every country around the world that made it available[2]. This is also available on GitHub and updated daily or as soon as new data comes up [143].

[1] https://github.com/pcm-dpc/COVID-19
[2] https://github.com/owid/covid-19-data

Most countries exhibited a very evident weekly oscillation in the number of new cases. This dynamic was due to the procedure behind testing and reporting in most countries and can be safely removed with a moving average when dealing with the time-series. While some put forward the idea that this periodicity is connected to weekend activities and gatherings, the hypothesis has been ignored as the behavior persisted even during periods of lockdown and restriction to people movements.

### 5.1.2 Tourism and Demography Data

In section 5.2 we used data on Italian domestic tourism together with other demographic indicators. These datasets come from the *Italian Institute of Statistics* (ISTAT) and can be easily and freely accessed on their website, which also allows the creation of custom views of the database to be downloaded in the most common formats.

At the time of our experiments, tourism data for 2020 were not yet available but we hypothesized that the numbers would be somewhat proportional to those of 2019 and thus used them as a proxy. This choice is not uncommon and a number of other studies followed a similar path. Another problem is the fact that the tourism numbers came aggregated monthly, whereas COVID-19 data is published daily. To circumvent this, we devised a clever aggregation strategy documented in Section 5.2.1.

ISTAT publishes data concerning both the number of people registering in hotels and other structures (arrivals) and the average number of days they stay (presences), aggregated from region of origin and region of destination. We used the former datum, as we were only interested in the volume of people moving from one region to the next. For each region thus we summed the number of incoming tourist from other regions with the number of people from that region that visited others. This is because we wanted to include the information about people leaving and then returning home with the virus in incubation.

Beside tourism data, we also included a number of demographic features, listed below, to control for the effect of other factor that could be contagion drivers. For those, the most updated information available at the time was used.

- Population density.
- Share of population over 65.
- Annual expenditure on healthcare for each region.
- NUTS-2 classification for each region (North-East, North-West, Center, South, Islands).

### 5.1.3 Other Data

#### School Re-openings Dates

In section 5.3 we used a changepoint detection method to compare the sudden acceleration in the spread of the virus in September 2020 with the reopening of schools in every Italian region. The dates are the official ones decided by the Italian Ministry for Education (MIUR).

#### Euro2020 Data

In section 5.4 we used the same method and compared our results to the dates of each match for the 2020 European Football Championship (EURO2020) which, being an anniversary edition, took place in many different countries instead of a single one. The official dates and location come from the UEFA website dedicated to the competition.

#### Köppen Classification

In Section 5.5 we studied COVID-19 seasonality on different countries around the world. We wanted to select as many different climates as possible, following the Köppen climate classification [144]. This classification divides climates into five main groups, where each group is considered based on seasonal precipitation and temperatures. The five main groups are: Tropical (A), Dry (B), Temperate (C), Continental (D) and Polar (E).

We analyzed 30 different countries that cover all the five groups, with several of the selected countries belonging to two or more groups, given their vast geography (e.g., India, Russia, and the USA, to cite a few). The complete list of the 30 countries follows below, each with its prevalent type of climates: Argentina (B, C), Australia (A, B, C), Austria (D, E), Belgium (C), Brazil (A, C), Canada (C, D, E) Chile (B, C, D), Colombia (A, C), Croatia (C), Denmark (D), France (C), Germany (C, D), Hungary (D), India (A, B, C, D), Indonesia (A), Italy (B, C), Japan (A, C, D), Mexico (A, B, C), Morocco (B, C), Norway (D, E), Portugal (C), Russia (D, E), Saudi Arabia (B), South Africa (B, C), South Korea (C, D), Spain (B, C), Sweden (D, E), Turkey (B, C, D), UK (C) and USA (B, C, D, E).

Notice that our selection includes 18 out of the 20 countries of the Group of 20. China was excluded just because its SARS-COV-2 data are not made available on a regular basis. Also, the European Union (EU) was not considered as a whole. Yet, in place of EU, we included the following EU members: Austria, Belgium, Croatia, Denmark, France, Germany, Hungary, Italy, Portugal, Spain, and Sweden.

## 5.2 Domestic Tourism During Summer 2020

The SARS-CoV-2 virus that emerged in Wuhan, China, at the end of 2019, causing the COVID-19 pandemic, has spread globally extremely quickly. This is unquestionably due to its high infectiveness, but it would not have reached the planetary scale if not for the *interconnectedness* of the world we live in. The centrality of Wuhan and its multiple transport communication hubs (roadways, railways, airports, and boats) played an important role in the spread of the virus in China in the early days of the infection, showing that those regions that required the least travel time from that city were hit earlier and more fiercely [145]. From there the virus was carried by planes, and by the beginning of March 2020 it had reached other Asian countries, then Europe, Australia, and the Americas.

At that point, it was clear that we were heading towards a pandemic, and most countries in the world put in place multiple restrictions to avoid the further spread of the virus, including banning international travel (even domestic in some countries) and closing public places, schools, and offices, as well as encouraging strict personal hygiene rules to reduce the possibility of virus transmission. These measures were fortunately quite effective in flattening the curve, up to the point that many countries lowered their guard by the end of spring 2020, thinking that the worst was behind them, and ended up facing a second wave in the autumn.

As Tomas Pueyo discussed in his article in the New York Times, the decision to forbid travel, both international and domestic, was one of the key ingredients in slowing down the contagion and making it manageable [146]. He goes on to analyze how *The Fence*, that is, the set of containment measures, including travel restrictions and strict monitoring of incoming tourists, had been progressively dismantled with the arrival of summer, making the sacrifices made during the previous months almost useless, as it often took just a few careless travelers to ignite a contagion in a region that was previously unscathed.

Those considerations describe quite well what happened in Italy during the period from February to August 2020. Essentially, after the outbreak of COVID-19 that started at the end of February in Lombardy and then violently hit all of northern Italy, the Italian government imposed a nationwide lockdown (on 9 March) in order to flatten the curve and avoid further spreading of the virus [147]. This measure proved to be effective with the passage of time, and thus the lockdown was gradually lifted at the beginning of May 2020 as the number of new daily active cases steadily declined; all citizens were then allowed to go outside, even when not strictly necessary, and people could gather, for example, in bars and restaurants while still maintaining social distancing. Nonetheless, it was not possible to move freely between regions until 3 June 2020 , when the total number

of new infected nationwide was averaging between 200 and 300 [148]. The count of the new daily infections remained stable until the last week of August 2020, when there was a noticeable uptick that brought the number of new daily cases, on a national level, to over $1,200$ [149].

Considering that the virus is known to take 10–14 days to show its symptoms, this puts the start of this inflation in the first half of August, which is the preferred time of the year for vacation trips in Italy, thus leading to a hypothesis, which is at the center of a national debate, that there was a significant relationship between Italian domestic tourism and the resurgence of the virus during the summer of 2020 [150]. What should also be considered is the fact that, this summer, the most typical Italian tourist destinations were visited mostly by Italian tourists, with little or no contribution by international tourism because many countries had set limits on travel, as well as implementing flying restrictions [151]. This idea of a relationship between Italian domestic tourism and the resurgence of the virus is also reinforced by a careful observation of the number of new daily infections in certain Italian regions that are typical holiday destinations; they were almost virus-free at the beginning of the summer, and, in the span of a few weeks, cases escalated from almost 0 to hundreds, daily. A peculiar example of this phenomenon is the insular region of Sardinia, where the virus had almost disappeared from the island and then suddenly came back in August 2020, surely carried by the tourists that arrived for their summer holidays [152].

During the first year of the pandemic, travel and its role in the spread of COVID-19, has been researched intensively, from various angles. However, most works only analyzed the phenomenon in the context of international travel or domestic travel in other nations, while we wanted to look at the situation in Italy.

Some works focused on the dynamics of people moving from one place to another, while others studied the way the virus can be transmitted in the environment of trains and planes [153, 154]. Krisztin et al. used econometric models to study how cross-country air travel played an important role in the early spread of the virus between European countries [155].

Dasgupta and Wheeler created a model that tries to estimate how the number of infections evolved as a function of people's interaction and the COVID-19 infection rates, registered in different geographical areas [156]. This model integrates estimates about the movement of people from area to area generated using a technique known as gravity modeling. Zhang et al. used a kind of gravity model based on the number of infections, which performed a regression analysis of various means of transport in order to investigate how relevant their contribution to the spread of the virus was [145].

In the same vein as [155], Farzanegan et al. studied the role played by tourism in the spread of the virus between countries across the world by using regression analysis to verify the relevance of tourist flows on the total number of infected people in conjunction with other factors, such as, for example, population density and aging, healthcare expenditures, and others [157]. Similarly, Falk and Hangsten analyzed how alpine tourism during the winter holidays resulted in a surge of COVID-19 cases in the Scandinavian countries, as many citizens brought back the virus after their return from their holidays in Italy and Austria [158]. Along the same line, Gössling et al. have investigated the influence of tourists moving from China to other countries, while comparing the resulting COVID-19 spread with previous pandemics, also with a special focus on the impact it had on economies [159]. The regression approach is the one we decided to use for our analysis described later.

We conclude this section by reporting on two other papers that are inspired by different visions of this problem. One of these two belongs to the group of simulation studies that are more interested in finding out how the virus spreads or decays, with the precise goal of guiding policy makers, yet without a meticulous attention to the precision with which they predict how many people are getting infected. To this group belongs the study conducted by D'Orazio et al., who investigated how certain COVID-19 containment measures could be more effective than others in limiting the spread of the virus in touristic cities, with the intent of going back to a business-as-usual regime, being pressured by economic reasons [160].

Finally, Susceptible-Infected-Removed (SIR) are the most popular modeling tools for epidemiology. These models are based on differential equations that describe the dynamics of the epidemic and need to calibrate their parameters on data in order to work correctly. Unfortunately, as Roda et al. have shown, there may not be enough data available, as the contagion unfolds, to calibrate them correctly and, more importantly, the more complex the model is (with an increasing number of parameters), the more this calibration becomes unmanageable [161].

### 5.2.1 Methods

Let us describe the methods we anticipated. We utilized a common change point detection method to find the most likely moment of change in the time series of the new daily infections [162]. Results show that the hypothesis of a relevant change in August 2020 had a solid statistical confirmation, which is valid for all the Italian regions included in this study, and it is centered in the last days of August 2020. We then performed a regression analysis using generalized linear model (GLM) estimating, on a monthly basis, the total amount of new infections

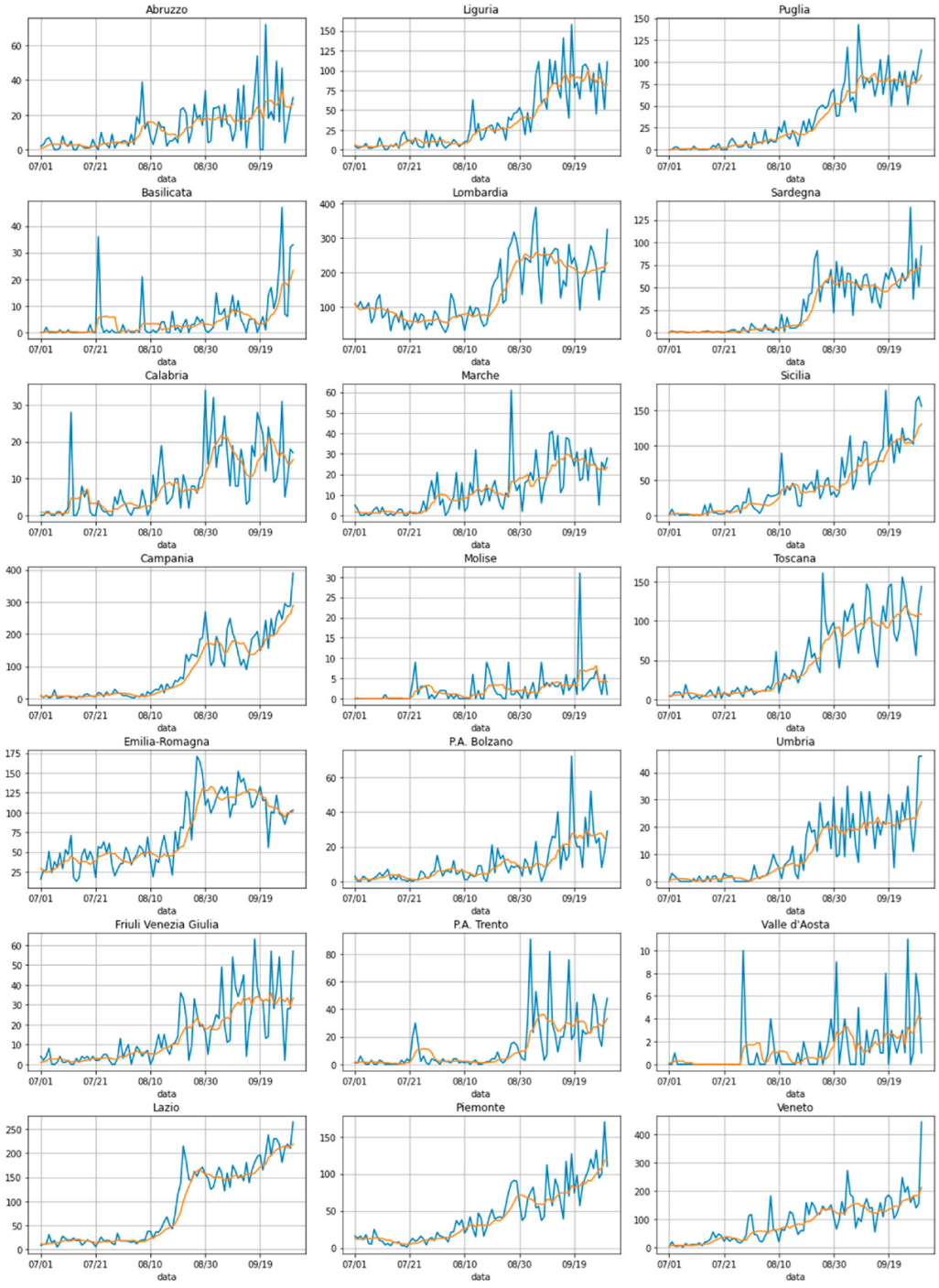



Figure 5.1: Number of new daily infection cases (in blue) and seven-day moving average (orange) for each of the 21 Italian regions, in the period between 1 July and 30 September.

using the number of inbound and outbound tourists per region along with a variety of other additional factors that could help explain the dynamics of the pandemic, beside tourism (e.g., regional population density, public sanitary expenditure per region). The end goal for this model is not predictive accuracy, although that is an indicator of a proper fit, but rather is to investigate contribution of certain features to the model performance and their statistical significance. In this, we can find a clue of their importance in the spread of the pandemic, hopefully guiding future policies.

### Changepoint Detection

The first thing we wanted to do was to check whether there was an actual change in the number of new cases compatible with our hypothesis that summer tourism could be the cause. This means that there should be a increase visible around 15 days after the contagion.

To find this change point in all the time series represented in all the 20 graphs of Figure 5.1, we used the change point selection method described in [163]. The idea is that of computing a discrepancy measure between different parts of the observed time series. If the two parts do not present a relevant change (for example, in their mean or in their variance, or both) this discrepancy factor will be low, otherwise there will be a peak, and a change point will be detected. This is essentially a window-based method, whose algorithm behind works by calculating a discrepancy measure d, based on the following formula:

$$d(y_{u..v}\ ,\ y_{v..w}\ ) = \ c(y_{u..w}\ ) - \ c(y_{u..v}\ ) - \ c(y_{v..w}\ ) \tag{5.1}$$

Where $y$, in our case, is a cumulative count of infected cases, while $u, v$ and $w$ represent those days that delimit the boundaries of two temporal contiguous windows. More precisely, with $u < v < w$, we have two temporal windows: the first one spans from day u to day v, the second one from day $v$ to day $w$. Obviously, these are two contiguous windows, hence if we reunite them, we yield a unique temporal window, spanning from day $u$ to day $w$. The cost function $c$ utilizes the square of the $L_2$ norm, and it is defined as follows:

$$c(y_I) = \sum_{t \in I} ||y_t - \bar{y}||_2^2 \tag{5.2}$$

where $I$ is the above-mentioned temporal window; $y$, again, is the cumulative count of the infected cases, while $\bar{y}$ is the correspondent mean value computed over that considered window. To individuate each change point for each of the Italian regions under observation, we looked for the maximum value taken by $d$.

A method able to find more than one changepoint (or none) by using some kind of threshold or penalty can also be considered. Using a penalty based on Bayesian Information Criterion (BIC), as suggested in [163], we computed a variable number of changepoints with the same window-based method that are indicated in purple in Figure 5.4.

### Generalized Linear Model

The method we decided to use followed the example of by Farzanegan et al. and consists in the analysis of a regression model that includes parameters for tourism and other relevant factors for control purposes. We fit a generalized linear model (GLM), assuming a negative binomial distribution for the target variable, estimating the parameters with maximum likelihood estimation. The software is implemented in the R programming language. The choice of a negative binomial distribution comes from the fact that we are dealing with counting data [157, 164, 165]. A possible alternative would be to use a Poisson distribution, which assumes equal mean and variance, making it unsuitable in the context of COVID-19 due to the well-recognized characteristics of over-dispersion manifested by the spread of this kind of disease, often referred to as a super spreading scheme [166, 167].

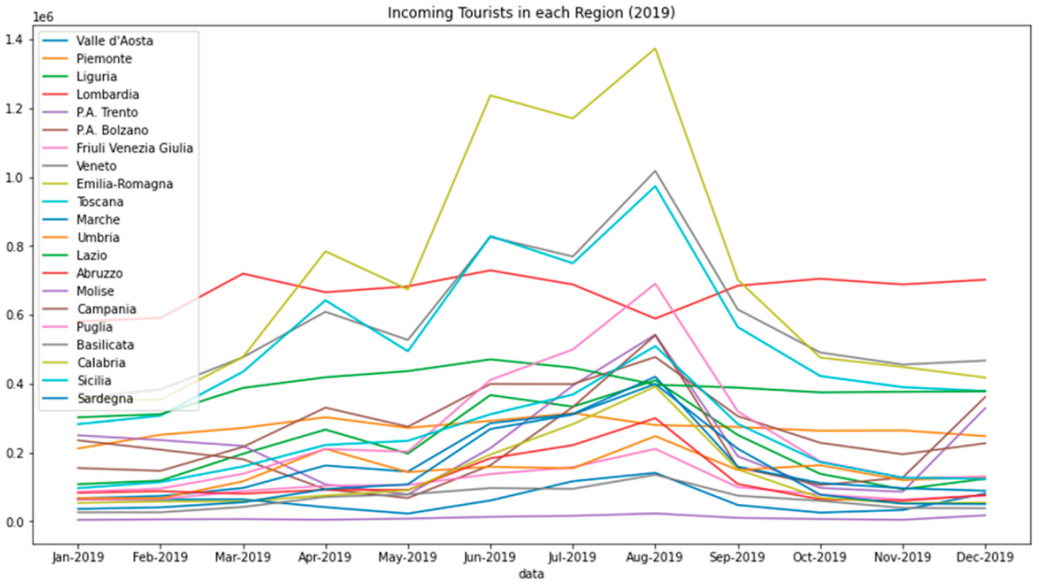



Figure 5.2: Incoming domestic tourists per each Italian region for each month of 2019. Typically, peaks of the curves are observed in August.

The target variable is the number of new cases for each Italian region in the period we are considering, shown in figure 5.1. Even though the COVID-19

data is released daily, we aggregated it on a monthly basis in order the be able to use data about tourism flows. The number for incoming tourist are in fact aggregated monthly, as shown in Figure 5.2, and there are no other sources with finer granularity.

To increase the number of observations in our dataset, we ended up creating a series of windows move by 1-week steps, as shown in Figure 5.3. For each "step" in-between months we took a fraction from the current month a fraction from the other, essentially obtaining a linear interpolation between the two values.

As an example, consider a one-month long window starting at the beginning of the second week of July and ending after the first week of August; this returns a total number of tourists to be taken into consideration, which is equal to the sum of 75% of the number of tourists for July plus 25% of the number of tourists for August.

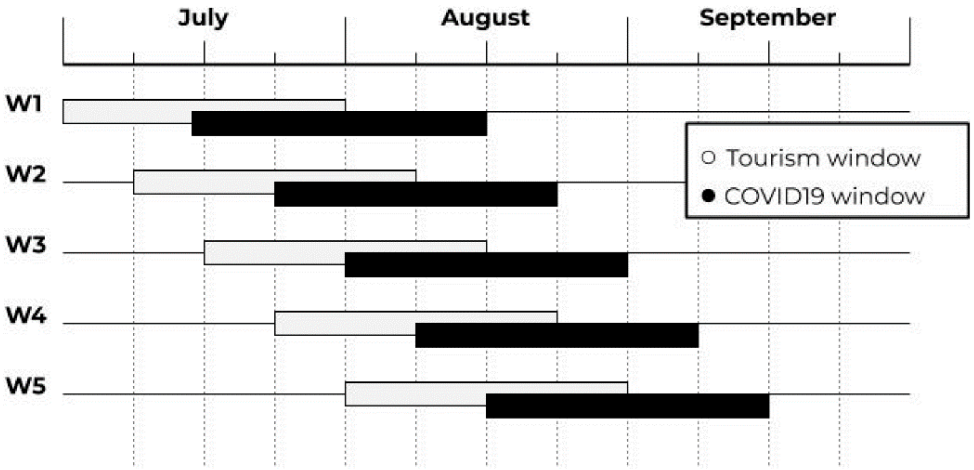



Figure 5.3: Dataset structure. Each window for the target (black) is shifted 14 days forward from the inputs (white) to account for the time required by COVID-19 symptoms to manifest.

As anticipated, we also included some feature to control for other relevant factors like population density, healthcare expenditure, aging population and geography. The formula for our negative binomial regression model is given below:

$$ln(\mathbf{y}) = \beta_0 + \beta_1 ln(T_{iw}) + \beta_2 ln(D_i) + \beta_3 ln(H_i) + \beta_4 O_i + \beta_5 A_i \tag{5.3}$$

with $\mathbf{y} = E(Y_{iw}|T_{iw}, D_i, H_i, O_i, A_i)$.

The various $\beta$ indicate the coefficients to be estimated, associated with the following independent variables:

- $Y_{iw}$ is the cumulative number of the new infection cases occurring in a region $i$ during the time comprised within the (black) window $w$;
- $T_{iw}$ is the sum of inbound and outbound tourists for a given region, $i$, during the aforementioned window, $w$;
- $D_i$ is the population density for each region, $i$, measured as the number of inhabitants per $km^2$;
- $H_i$ is the healthcare expenditure for the region, $i$, expressed as a percentage of the region's GDP;
- $O_i$ is the proportion of total the population over the age of 65;
- $A_i$ is a categorical variable that indicates the NUTS2 classification mentioned in Section 5.1;
- $ln$ stands for the natural logarithm. We applied this transformation to be comparable with the approach in [157] and also because it showed improved performance.

### 5.2.2 Results

Let us start by considering Figure 5.4, which shows the main changepoint in red and, in purple, the occasional others resulting from the threshold method described before. The most likely change point for majority of the Italian regions falls around the end of August, placing the peak of the COVID-19 spread around the middle of that month (owing to the well-known 14-day lag). This period in time is exactly when most Italians usually go on holiday. There are, however, a few exceptions, namely Abruzzo, Basilicata, Molise, and the autonomous province of Bolzano, whose change point is found to be closer to the middle of September. Nonetheless, it is important to note that those regions, being small in size, typically receive flows of tourists more gradually, never exceeding large volumes. Hence, the reason for that shift in time could lay there. Because of its nature, the method also has a harder time when called on to detect changes if they occur slowly and gradually, such as in the case of Sicily or Marche.

Moving on to the regression model, we begin by discussing the estimated coefficients. While our regression model can be used as a prediction model, as we show further down in this section, the real interest of regression analysis is inference about the factors included in the model, looking for statistically significant relationships between them and the target variables.


Figure 5.4: Number of new daily infection cases (in blue) and correspondent 7-day moving average (orange) for each of the 21 Italian regions, in the time span between July 1st and September 21st. In red marked is the change point found by the change point detection method. In purple are showed additional changepoints.

Table 5.1: Coefficient Estimates for our generalized linear model (GLM). They show how tourism ($T$) and density ($D$) are highly significant, as well as the geographical indication ($A$) for Island and Southern regions (although less). The percentage of elderly ($O$) is also included.

| **Coefficients** | **Estimate** | **Std. Error** | **z-Value** | **Pr($>|z|$)** |
|---|---|---|---|---|
| (Intercept) | -9.93064 | 2.10982 | -4.707 | 2.52e-06 |
| log(T) | 0.85395 | 0.12426 | 6.872 | 6.31e-12 |
| log(D) | 0.82921 | 0.14986 | 5.533 | 3.14e-08 |
| log(H) | -1.19588 | 0.59292 | -2.017 | 0.04370 |
| O | 11.12581 | 3.68649 | 3.018 | 0.00254 |
| (A) Islands | 1.33754 | 0.35277 | 3.792 | 0.00015 |
| (A) North-Est | 0.06887 | 0.17690 | 0.389 | 0.69705 |
| (A) North-West | -0.31216 | 0.19199 | -1.626 | 0.10397 |
| (A) South | 0.65365 | 0.27646 | 2.364 | 0.01806 |

Table 5.1 contains the estimates produced by the fitted GLM and their levels of significance. Under Coefficients, the reader can find the names of the predictors whose coefficients need to be estimated. Underestimates, one can find the estimates of those coefficients, as per Equation 5.3. Along with those estimates comes an attempt to measure the errors made by the model when those estimates are computed. More precisely, the third column (Std Error) presents the standard deviations associated with those estimates, while the fourth column provides the so-called $z$-value, which is a statistic that returns a result different from the hypothesis, which is that the given coefficient is equal to zero. Put simply, when $z$ is high (either positive or negative), there is a low probability that the coefficient under consideration is zero; that is, not relevant for the solution of the regression problem of interest. Finally, we come to the fifth column, where the probability that the estimates set under the column Coefficients can exceed the modulus of the already explained z-value is estimated. This last value, in particular, should be considered for an immediate analysis of the table, where the lower values under column 5 are associated with the most relevant predictors.

All this being said, what is quite interesting in this discussion is the careful observation of the estimates computed for the coefficients of the following predictors: tourism and density of population ($T$ and $D$). These two factors appear as important predictors with the capability to have an influence on the solution of the regression problem we are facing, as confirmed by their high significance, given that the estimates of their relative coefficients come with good precision, as an analysis of columns 3, 4, and especially 5, reveal.

It is also interesting to note how the Island and Southern regions seem to be linked to an increased number of infections (with respect to other Italian regions) because of the same statistical motivations we have explained before. Also of note, for similar reasons, is the predictor associated with the percentage of persons aged over 65 ($O$), while other predictors (e.g., healthcare expenditure, $H$) do not seem to play a relevant role for this model.

In summary, this model appears to be more influenced by how many infections are brought by tourists; the percentage of people with an age over 65 is of secondary importance, followed by how much money a given region spends on healthcare.

### Predictive Performance

Even though it was not in our interest to create a predictive model we wanted to see how this regression model would fare against a deep neural network, albeit a very simple one, trained on the same data.

| Region | Real Cases | GLM | Neural Net |
|---|---|---|---|
| Abruzzo | 480 | 463 | 293 |
| Basilicata | 135 | 91 | 117 |
| Calabria | 383 | 298 | 270 |
| Campania | 3959 | 1466 | 2007 |
| Emilia-Romagna | 3325 | 2133 | 2484 |
| Friuli Venezia Giulia | 677 | 362 | 348 |
| Lazio | 4303 | 1619 | 2168 |
| Liguria | 1504 | 907 | 889 |
| Lombardia | 6239 | 3349 | 4329 |
| Marche | 542 | 513 | 470 |
| Molise | 84 | 37 | 101 |
| Piemonte | 1817 | 781 | 983 |
| Puglia | 1689 | 922 | 857 |
| Sardegna | 1452 | 491 | 610 |
| Sicilia | 1625 | 1254 | 1034 |
| Toscana | 2412 | 1314 | 1557 |
| Trentino-Alto Adige | 869 | 472 | 778 |
| Umbria | 546 | 223 | 199 |
| Valle d'Aosta | 46 | 40 | 67 |
| Veneto | 3852 | 2252 | 2557 |

Table 5.2: Cumulative number of new infection cases as predicted by different models for the interval 15 August – 15 September 2020.

Table 5.2 shows the results of this prediction test. As we can see the neural network is slightly better but not by much while also being an opaque model that cannot be interpreted in any way. In this situation traditional methods are to be preferred since the benefit of interpretability is of greater value than the small loss of accuracy. If we look at Mean Absolute Percentage Error, how far from the actual number each model is on average, the neural networks scores 36% and the GLM scores 40%, only a 4% difference.

## 5.3 School Reopening in September 2020

As the summer of 2020 was coming to its end, with its relatively low number of COVID-19 cases in most western countries, many started arguing whether it was wise to normally restart school activities. During the first wave, in most nations schools were closed, as any other activity, and only some partially reopened them as the situation got under control and the lockdown was lifted. In this context, the large COVID-19 outbreak in a high school in Israel in May 2020 ignited the discussions about the role of schools in the spread of the virus [168, 169].

Two opposing sides appeared quite clearly: on one side, those who considered schools to be a minor risk and the importance of school paramount; on the other, those who were concerned by the lack of clear data on the contagion dynamics in schools and were scared by the high number of asymptomatic cases in younger people. The same discussion emerged in all the other countries affected by the ongoing pandemic, with the two sides bringing to the table mostly the same arguments with the occasional country-specific remarks. Most international literature that suggests the absence of considerable risk factors connected to schools focuses on the fact that children and adolescents seem to be the least affected by the virus, both in terms of the number of positive cases but also of symptoms and contagiousness. Ismail et al,[170] studying UK schools, show how the incidence in students is not larger than the total incidence in the region and how the most cases inside schools are transmission between staff members. Similar studies targeting other countries draw similar conclusions, especially when they come to primary schools and kindergarten [171–176]. The hypothesis that asymptomatic children could easily and unknowingly spread the virus in their families was rejected by Munro and Faust [177] and by other similar analyses, based on argument that children have a lower susceptibility to the virus and thus play a lesser role in the transmission [178–180].

The researchers presenting the opposite hypothesis point to the weaknesses of many of the aforementioned studies, namely the very small samples and the fact that often the role of asymptomatic subjects is not tracked and considered properly. Many point out the correspondence between the insurgence of the second wave in many countries within 2–3 weeks from the school openings and point to the data that suggest a higher spread in the school-aged population in those months, especially in high school and university students [181, 182]. Flasche and Edmunds have responded to Ismail's study saying that it was conducted with schools not fully populated underestimating the potential of children, especially in the age bracket 10–18 years. [183] This group has seen a considerable increase in September, as did college students, and seemed to be a common source of SARS-CoV-2 infections in the households. Also, Yamey and Walensky [184] express

their concerns for universities reopening, while Sebastiani and Palù [185] studied the Italian situation and argued that the rise of new cases in September, with most SARS-CoV-2 infections happening in the household, was compatible with the hypothesis of school being a factor. While inside schools, measures were taken, they argued that outside contact was inevitable due to public transportation and social gatherings, so that young people spread the virus among themselves. Larosa et al. [186] conducted a study in the Reggio Emilia province (Emilia-Romagna region, Italy) showing that there were non-negligible clusters in the age bracket $10 - 18$ years. They also suggested that more prompt isolation and testing could have hindered the spread, stressing how important timeliness is in this context. Despite their opposing positions, most researchers emphasize the need for the same measures: an active case finding approach with systematic and thorough testing of students and personnel.

Following this scientific debate, we focused on Italy and looked at the contagion curves and relating them to the dates schools opened in each of the 21 regions. Italy has faced a hard time during its second wave, as it was bringing the healthcare system to its knees. It started during the autumn, somewhere around the start of October 2020, and peaked in November when the government imposed a new form of lockdown with color-coded regions depending on risk.

September was a crucial period, as with the end of summer many activities were going back to normal, and the virus prevalence in the nation was quite low [187]. In the first days of that month, a slight increase in the number of new cases was registered in most regions, probably due to the cross-regional movement for the summer holidays [137]. This prompted many to warn of the arrival of the second wave, but the number of new cases stabilised in the coming days, and the growth was considered small in any case. School reopened between the third and the fourth week of September [188, 189] while people had also already started going back to their offices and activities. Another thing to notice is the referendum of 20 September that in some regions corresponded with other elections for local government and senate representatives. While attendance was quite low, one could wonder the effects of such an event.

Because of the shocking scarcity of available and reliable data on SARS-CoV-2 infections inside Italian schools, we chose another perspective in order to investigate the hypothesis of an association between schools and the resurgence of the virus, by analysing the growth rate of the total number of infections in all the Italian regions, before and after the school reopening.

### 5.3.1 Methods

Given the scarce availability of data collected in schools that could better describe their role in Italy (for the reasons stated in the Introduction), we decided to work with the population-wide data at the regional level. We fitted a piecewise linear regression model where the dependent variable was only the number of new daily confirmed COVID-19 cases, and the independent variable was just the number of days since 1 September 2020 (until 31 October 2020). The result is a model comprised of a changepoint and two segments, whose slopes represent the growth rate before and after the acceleration in the exponential growth causing the second wave; we passed from a stable situation with the exponent close to 0 which means little to no growth, to a situation with markedly positive exponent. To have a measurement of the uncertainty in our estimates, we decided to use a Bayesian framework for the regression as described in [190]. Two transformations were applied to the initial data. First, we used a 7-day rolling average as the raw data presents a weekly periodicity due to the way COVID-19 tests are carried out and registered. Second, we applied a natural logarithm so that the exponential growth appears as an easily identifiable slope. Using the R package called `mcp` built on top of JAGS [191], we estimated a piecewise linear regression model for each region. We modeled our dependent variable $ln(y)$ (the natural log of confirmed daily cases) as a Normal distribution whose mean depends on the regression coefficients $a_1$ and $b_1$ (the intercept and angular coefficient) before a changepoint $\tau$, and $a_2$ and $b_2$ after the changepoint, as represented in the formula below.

$$\begin{aligned} ln(y) &\sim \mathcal{N}(a_1 + \mathbf{x}b_1, \sigma^2) \quad \text{if} \quad x \leq \tau \\ ln(y) &\sim \mathcal{N}(a_2 + \mathbf{x}b_2, \sigma^2) \quad \text{if} \quad x > \tau \end{aligned} \tag{5.4}$$

The model was fitted using a Markov Chain Monte Carlo (MCMC) method. Since the two lines are joined at the changepoint, without discontinuities, the second intercept term $a_2$ is not estimated as is bound to be $a_2 = \tau(b_1 - b_2) + a_1$.

As described in [191], the intercept and slope priors used for the Bayesian estimation were chosen in Gaussian families, while for the changepoint a Uniform distribution was used, precisely as reported below:

$$\begin{aligned} \tau &\sim \textbf{Uniform}(\text{min(x)},\ max(x)\ ) \\ \sigma &\sim \mathcal{N}(0,\ sd(y)) \\ b &\sim \mathbf{t}(0, 3 * sd(y),\ 3) \\ a &\sim \mathbf{t}(0,\ sd(y)/max(x) - min(x)) \end{aligned} \tag{5.5}$$

The final estimates for all the parameters are reported in Table 5.3 with the 95% CI. To give an idea of the increase in slope, we computed the number of days

($DT_1$ and $DT_2$) necessary to observe a doubling in the number of new cases from the changepoint onward, with both slopes (using the average angular coefficient), as shown in the following equations:

$$DT_1 = \frac{ln(2y_\tau) - a_1)}{b_1} - \tau;\ DT_2 = \frac{ln(2y_\tau - ln(y_\tau))}{b_2}; \quad \text{where } y_\tau = e^{a_1+\tau b_1} \tag{5.6}$$

### 5.3.2 Results

Table 5.3 shows all the estimated coefficients for each region. Similarly, Figure 5.5 shows all the curves and regression lines. Out of the 21 Italian regions, 15 (71%) of them have a changepoint within 28 days from the date when the school opened. In particular, the average number of days between the opening and the changepoint is 16.66 days (CI 95% 14.47 to 18.73). This number is plausible in a scenario where one has to be exposed to the virus and then manifest symptoms in order to be tested, also considering that children and adolescents were often asymptomatic and could have functioned as a driver for "second-degree" contagion.

Looking at the estimated slopes of Figure 5.5 and 5.6, we can have an idea of the strength of the increase. While one could comment on the difference in the angular coefficients, converting it into an angle, the number does not convey the idea very effectively. Translating it into the number of days required for a doubling in the growth rate is more interpretable.

Of the 15 regions mentioned above, 4 had a slope that is null or slightly negative before the change, making this time infinite, but the fact that the trend was inverted to an average value of 6.23 days (CI 95% 4.30 to 8.20) is significant enough by itself. The remaining 11 regions went from an average of 47.50 (CI 95% 37.18 to 57.61) days at the rate before the changepoint to an average of 7.72 (CI 95% 7.00 to 8.48) at the rate after the changepoint. Figure 5.5 illustrate these results.

Of the 6 regions that break the pattern, two of them, P.A. Trento and P.A. Bolzano (often considered one region, called Trentino-Alto Adige) presented a changepoint more than four weeks after the school opening; they are indicated in blue in Table 5.3. On the other hand, the remaining four, in yellow in Table 5.3, begin their rapid increase before or in correspondence of the school opening. Both are displayed in Figure 5.6. All the models converge quite nicely, with narrow 95% CI, except for Basilicata, where the changepoint estimation is extremely wide, probably due to the high variability in the reported numbers.

Figure 5.5 and 5.6 use the following graphical convention: The vertical blue line indicates the changepoint with the blue area corresponding to the 95% CI.

Table 5.3: Estimated parameters from the 21 Italian regions along with school opening date (Open), number of days between opening and changepoint dates (D) and the doubling time for the two slopes ($DT_i$). Between brackets are the 95% CI.

| **Region** | **Open** | **D** | $\Delta$ | $m_1$ | $m_2$ | $a_1$ | $\sigma$ |
|---|---|---|---|---|---|---|---|
| Basilicata | 23/09 | -4 | 19.25 (0.01, 51.02) | 0.02 (-0.00, 0.06) | 0.06 (0.04, 0.08) | 1.15 (0.78, 1.45) | 0.35 (0.29, 0.29) |
| Campania | 23/09 | -2 | 20.85 (18.81, 2.86) | -0.00 (-0.01, 0.01) | 0.07 (0.07, 0.08) | 5.07 (5.00, 5.15) | 0.08 (0.07, 0.07) |
| Abruzzo | 23/09 | 0 | 22.54 (20.73, 24.28) | -0.01 (-0.02, -0.00) | 0.10 (0.09, 0.11) | 2.73 (2.58, 2.88) | 0.18 (0.15, 0.15) |
| Sardegna | 21/09 | 0 | 21.26 (17.65, 26.73) | -0.00 (-0.01, 0.01) | 0.05 (0.04, 0.05) | 3.94 (3.85, 4.03) | 0.08 (0.07, 0.07) |
| Puglia | 23/09 | 8 | 30.92 (29.63, 32.20) | 0.01 (0.01, 0.01) | 0.07 (0.07, 0.07) | 4.13 (4.08, 4.19) | 0.08 (0.07, 0.07) |
| Calabria | 23/09 | 9 | 31.67 (29.68, 33.65) | -0.01 (-0.01, 0.00) | 0.10 (0.09, 0.11) | 2.75 (2.62, 2.88) | 0.19 ( 0.16, 0.16) |
| Valle d'Aosta | 13/09 | 13 | 26.22 (24.09, 28.29) | -0.00 (-0.02, 0.01) | 0.14 (0.13, 0.15) | 0.68 (0.49, 0.87) | 0.27 (0.22, 0.22) |
| Umbria | 13/09 | 14 | 26.87 (25.72, 28.09) | 0.01 (0.00, 0.01) | 0.10 (0.10, 0.10) | 2.79 (2.72, 2.87) | 0.11 (0.09, 0.09) |
| FVG | 15/09 | 16 | 30.91 (29.49, 32.22) | 0.01 (0.01, 0.02) | 0.09 (0.08, 0.09) | 2.96 (2.88, 3.05) | 0.13 (0.11, 0.11) |
| Piemonte | 13/09 | 16 | 28.58 (27.40, 29.73) | 0.02 (0.02, 0.02) | 0.11 (0.10, 0.11) | 3.99 (3.94, 4.05) | 0.08 (0.06, 0.06) |
| Veneto | 13/09 | 16 | 29.20 (27.09, 31.62) | 0.01 (0.01, 0.02) | 0.08 (0.07, 0.08) | 4.75 (4.68, 4.82) | 0.09 (0.07, 0.07) |
| Lombardia | 13/09 | 17 | 30.30 (29.56, 31.02) | -0.01 (-0.01, -0.00) | 0.12 (0.12, 0.13) | 5.49 (5.44, 5.55) | 0.08 (0.07, 0.07) |
| Toscana | 13/09 | 17 | 29.67 (28.81, 30.51) | 0.01 (0.01, 0.01) | 0.10 (0.10, 0.10) | 4.39 (4.33, 4.45) | 0.08 (0.07, 0.07) |
| Emilia Romagna | 13/09 | 18 | 30.65 (30.11, 31.23) | -0.01 (-0.01, -0.01) | 0.09 (0.09, 0.09) | 4.86 (4.83, 4.90) | 0.05 (0.04, 0.04) |
| Marche | 13/09 | 19 | 32.35 (30.53, 34.01) | 0.02 (0.01, 0.02) | 0.10 (0.09, 0.11) | 2.73 (2.63, 2.83) | 0.15 (0.12, 0.12) |
| Liguria | 13/09 | 20 | 32.74 (30.35, 35.16) | 0.02 (0.02, 0.03) | 0.08 (0.08, 0.09) | 3.85 (3.75, 3.96) | 0.15 (0.12, 0.12) |
| Molise | 13/09 | 22 | 35.16 (33.16, 37.35) | 0.02 (0.01, 0.03) | 0.12 (0.11, 0.14) | 0.72 (0.55, 0.91) | 0.29 (0.24, 0.24) |
| Sicilia | 13/09 | 22 | 34.52 (32.17, 36.81) | 0.04 (0.03, 0.04) | 0.07 (0.07, 0.08) | 3.81 (3.75, 3.87) | 0.09 (0.08, 0.08) |
| Lazio | 13/09 | 23 | 35.85 (34.69, 37.01) | 0.02 (0.01, 0.02) | 0.09 (0.09, 0.10) | 4.89 (4.85, 4.94) | 0.07 (0.05, 0.05) |
| P.A. Bolzano | 6/09 | 28 | 34.39 (29.46, 41.99) | 0.04 (0.03, 0.05) | 0.10 (0.07, 0.12) | 1.97 (1.83, 2.12) | 0.22 (0.17, 0.17) |
| P.A. Trento | 13/09 | 32 | 44.99 (42.97, 47.17) | 0.02 (0.02, 0.03) | 0.12 (0.09, 0.14) | 2.68 (2.55, 2.81) | 0.24 (0.20, 0.20) |

The yellow vertical line indicates the date school reopened. Black dots are the natural log of daily confirmed cases from 09/01/2020 to 31/10/2020 that we used as inputs. The regression lines, using the average value for the coefficients, are shown in red and green. On the rightmost side of each plot, the y-axis is reported without the log transformation, to allow the reader to infer what the confirmed case rates were in each region.

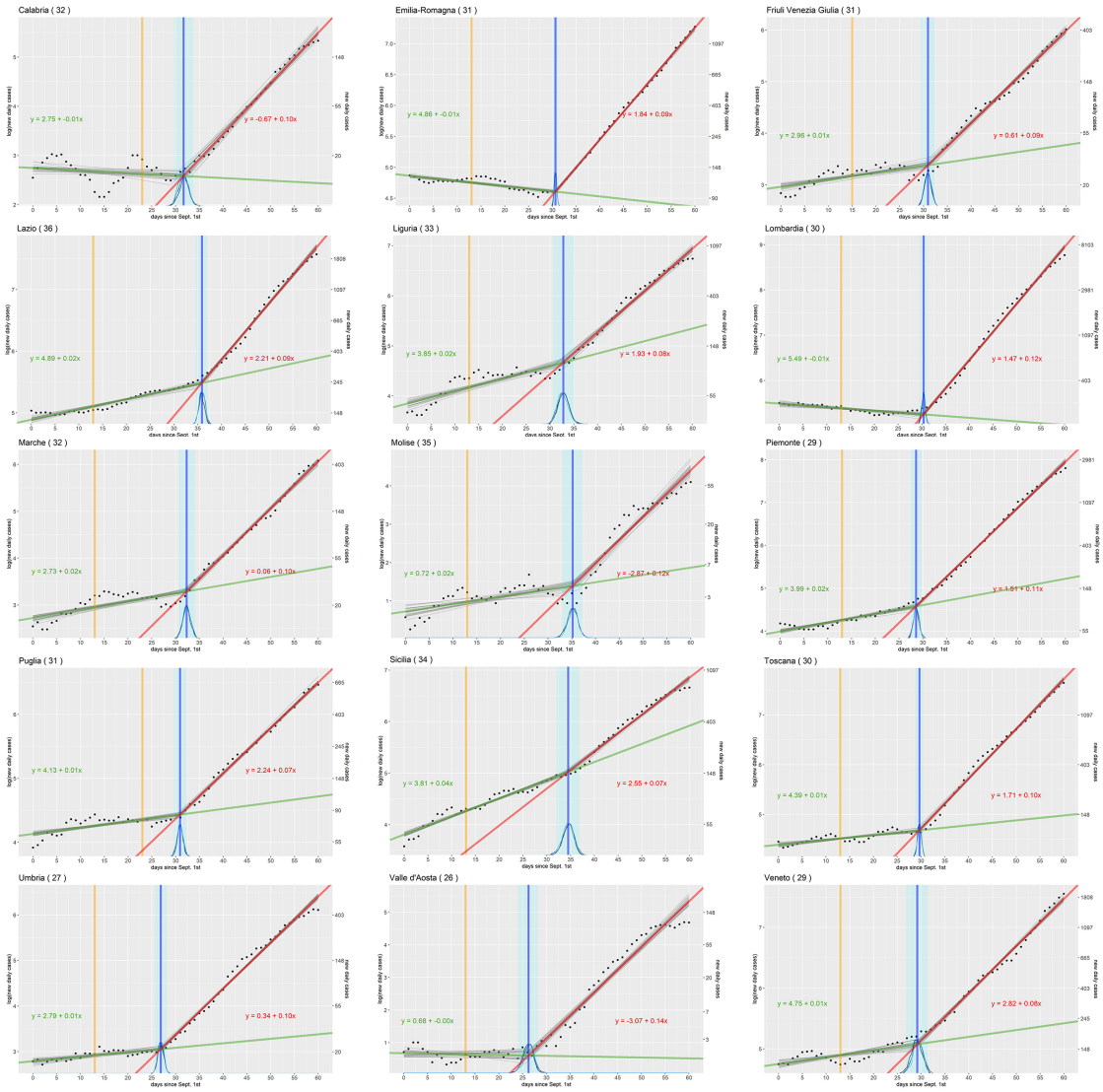



Figure 5.5: Plots for the regions whose changepoint is within 28 days since school openings.

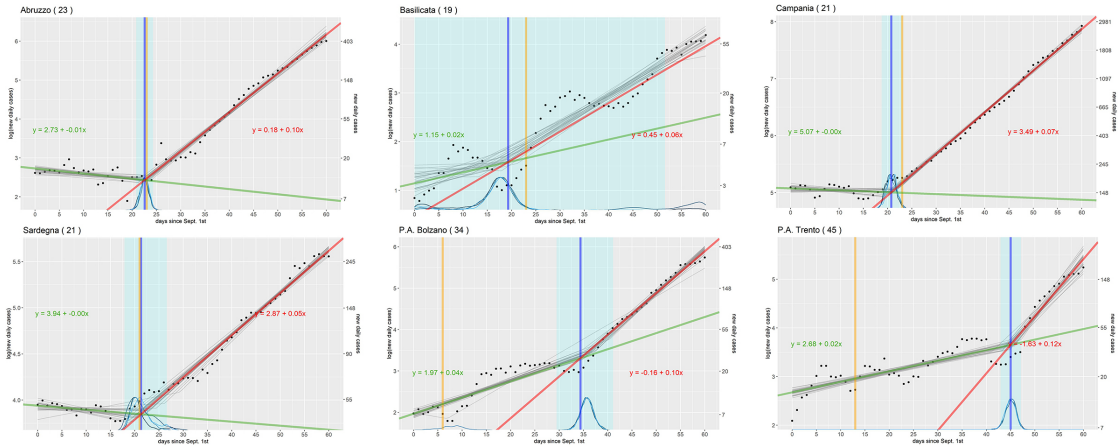



Figure 5.6: Regions that do not exhibit the pattern. First 4 start rising before school open, the last two take more than 28 days after that.

The results highlight how the second wave in Italy started in the days between September and October, with a degree of variability. These changepoints are on average a couple of weeks after the school openings. Certainly, multiple confounding factors played a role in the acceleration of the SARS-CoV-2 infection, but our opinion is that schools are surely one of those, and the magnitude of their effects should be investigated more thoroughly.

In the short period that precedes the second wave, there were not many events that interested as many people as schools. The referendum and elections on 20 September almost coincides with school opening but did not have a large participation. Workers going back to their offices could also have played a major role, however by looking at the nationwide mobility report by Google Mobility, we can see that in September the number of people moving to their workplace increased steadily from $-30\%$ to $-20\%$ with respect to the reference level before the pandemic [192]. If we consider that in Italy, there are approximately 25 million workers, an increase of 10% would translate into 2.5 million more people circulating. If we compare this number with that of students and school personnel which is equal to approximately 11 million, we can conclude that perhaps, even when considering that both categories use public transportation heavily, schools could be more influential in spreading the virus.

The regions that do not follow the pattern may tell us something more. In a regionalised country like Italy, with a strong territorial differentiation, Trento and Bolzano could be outliers [193]. They are two northern, neighbouring autonomous provinces, often considered a single region, which share characteristics of a higher care coverage of social and educational services that could set them apart from other regions. In the others, the change is before or coincidental with school

opening, so any factor that could have ignited the second wave is to be researched outside of the school activity and all the other connected activities. The effect of schools, if any, would be absorbed in the inflation already in act. Being all four maritime regions, tourism could perhaps be one of the major causes. Regarding those regions where the number of cases has high variability, the reason is most likely that the number of tests done each day varies as much. We could have normalised the cases with the number of tests, but this datum is often unreliable and leads to unrealistic normalised values, so we decided to avoid this. It should be also noted that any research hypothesis concerning SARS-CoV-2 infections in schools has a hard time being verified in Italy, as no region has so far seriously investigated the dynamics of the spread of the virus inside schools by using a systematic active case finding methodology.

There are, finally, some important technical considerations to put in evidence concerning the possible limitations of this study. First, the data we have used to count the number of daily SARS-CoV-2 cases were made available by the Italian government under the form of aggregated measures. In several cases, those measures have changed meaning/value over time, with errors that were never corrected. Not only that, but in order to highlight the shift in the infections growth rate (i.e., the slopes), along with the moment in time when this happened (i.e., the changepoint), we adopted a simple normal model, resulting in two lines before and after the changepoint of interest. We recognise that this method is not the more accurate one with which to count COVID-19 cases. Poisson-like distributions would be more appropriate. Moreover, it should be considered that the target of this study was not trying to create a model for the count of COVID-19 infections, per each single day of observation, but to look at how quickly they grew, comparing the slopes. Nonetheless, a Poisson model yields comparable results. In particular, the 15 regions that showed the pattern continue doing so (the changepoint falls within 3 weeks from the school reopening). Also, the relevant parameters (the slopes coefficients $b_1$ and $b_2$) for all the 21 Italian regions have an average absolute difference from their respective of less than 0.01 (0.0057 for $b_1$, 0.0081 for $b_2$), with the average number of days since the reopening of schools equal to 15.2 days, well within the CI computed with the normal model. A limitation of this study is also concerned with the impossibility to provide an estimate of many other confounding factors that could have played a role during the period of observation, besides schools. Another limitation resides in the use of Italian data. The extension to different geographies could result into more robust results.

## 5.4 2021 European Football Championships

During the summer of 2021 Europe, as well as other countries around the world, saw a resurgence in the COVID-19 pandemic, after a brief respite given by the effects of the vaccination that started in the first half the year. This new wave of the pandemic seemed to be driven by a new strain of virus that has been referred to as the Delta variant. This is the scenario in which the European football championship has taken place, from 11 June to 11 July 2021 (one year later than it should have been). This 2020 edition, being a special celebration for the 60th anniversary of the tournament, has had the peculiarity of being hosted by several different countries, instead of just one as it normally happens.

The decision to allow such a massive event across the European continent, in such a delicate time, immediately triggered a debate on the problems it would cause. Nonetheless, the competition was held, leaving each hosting country some freedom on which restrictions to apply (e.g., the number of fans allowed at each football stadium). This resulted in very different behaviors, ranging from Hungary hosting its matches at full stadium capacity at Puskás Arena ($\sim 68,000$ seats) to Germany limiting the attendance to 22% of the maximum stadium capacity [194–197]. Obviously, there were more factors than just the stadium, with fans, massively gathering in pubs, squares, and public places, to watch the matches, thus leading to infection clusters that surged all around Europe, as witnessed by the media coverage of these events [198–201]. Not only that but even the gathering of teams and their staff may have given their contribution to the spread of the virus (given the itinerant nature of this edition), as the COVID-19 literature on football and other sports suggests [202–204].

On one side, one could conclude that those who considered this event to be a minor risk did not take into any consideration of those theories that maintain that, with COVID-19, super-spreading events may be the main driver of an epidemic spread, under specific circumstances [205, 206]. An example, on 19 February 2020, was the Champions League match, between Atalanta and Valencia, which attracted a third of Bergamo's population to AC Milan's San Siro stadium. In addition, more than two thousand and a half of Spanish supporters took part. Experts, now, point to that 2020 football match as one of most relevant reasons why the city of Bergamo had become the epicenter of the COVID-19 pandemic, during the first wave in Italy, with a very high death toll; not to mention, that the 35% of Valencia's team also became infected [207].

On the other hand, it is well known that the return of supporters to stadiums is the highest priority for football's business, and the financial impact of the COVID-19 pandemic on football depends, almost exclusively, on both the timing and the

scale of supporters' return to stadiums [208].

Following this debate, this work focused on the European football championship and its matches, looking for a possible compatibility with the reversal of the decrease/increase trend of the SARS-COV-2 cases, observed in many countries participating in the tournament. To investigate the hypothesis of an association between those football matches and the resurgence of the virus, we searched for a changepoint in the daily timeseries of the new SARS-COV-2 cases registered in each country, expecting it to appear not later than two to three weeks after the date of the first match that the national team played.

Upon finding such a changepoint, we investigated if that changepoint was coincidental with a change in the infection rate, from a decreasing trend to an increasing one. It should be noted that our type of analysis has been observational in nature, and it was used to determine if the exposure to the specific risk factor, given the frequent mass gatherings following the football events, might have correlated with the particular outcome of the virus resurgence in many European countries. With this type of study, we cannot demonstrate any cause and effect, but we can make preliminary inferences on the correlation between the participation in the European football championship of a given country and the inversion in the SARS-COV-2 case rate that may have hit, at a particular point in time, the population living in that country.

### 5.4.1 Methods

The methods we used in this experiment are the same described in section 5.3.1. The only exception is the inclusion of a formula to also compute the halving time for decreasing time-series as observed for many countries in the months before the EURO2020 matches.

The following equation are for the halving or doubling time before the changepoint $\tau$:

$$H_b = \frac{\ln \frac{y_\tau}{2} - a_1}{b_1} - \tau; \quad D_b = \frac{\ln 2y_\tau - a_1}{b_1} - \tau \text{ where } y_\tau = e^{a_1 + \tau b_1} \tag{5.7}$$

While the following equation are for the halving or doubling time after the changepoint $\tau$.

$$H_a = \frac{\ln \frac{y_\tau}{2} - \ln y_\tau}{b_2}\ ; \quad D_a = \frac{\ln 2y_\tau - \ln y_\tau}{b_2} \text{ where } y_\tau = e^{a_1 + \tau b_1} \tag{5.8}$$

As a final note, it is important to mention that while it is quite usual that COVID-19 cases can show their biggest single-day jumps two to three weeks after a

particular mass event [185], we have extended the search space for a changepoint of our procedure to four weeks, for the sake of the reliability. Nonetheless, following the literature we have considered of interest only those changes occurred in the infection curves in the interval 5-6 to 22-23 days since the event of interest.

### 5.4.2 Results

This Section is split over two different parts. The first reports the results we obtained with the 22 countries that took part in the European Football Championship. The second illustrates the results we got with some 12 European countries that did not participate in the tournament.

#### Countries That Participated in the Tournament

17 out of 22 (77%) countries taking part in the European football championship have shown a changepoint occurring not later than two to three weeks after their first match (i.e., during the tournament). For all these 17 countries, the changepoint coincides with a reversal in the new daily SARS-COV-2 cases from a decreasing to an increasing rate.

The group of all these countries provides evidence in favor of the hypothesis. Precisely, the group is comprised of all the following countries: Austria, Belgium, Croatia, Czechia, Denmark, Finland. France, Germany, Hungary, Italy, Netherlands, North Macedonia, Poland, Slovakia, Spain, Switzerland, and Ukraine.

Table 5.4 provides the lists of those countries, where under the $\tau$ column we listed for each country the mean value of the days passed before the changepoint was detected, since the beginning of the period of observation (28 May 2021). Since we are working with a posterior distribution, in brackets the 95% CI are indicated. In the *Diff* column, instead, we have listed the difference, in terms of days, between the point in time when the changepoint occurred and the date of the first match played by that given national team. The fourth and fifth columns of Table 5.4 show the mean values (with the corresponding 95% CI) for the coefficients $b_1$ and $b_2$, that have been used to compute the steepness of the slopes, respectively before and after the changepoint. The sixth column, finally, reports the average value of the first intercept $a_1$, with its 95% CI.

We have further worked with the numbers comprised in Table 5.4 by rounding the mean changepoint value for all the 17 countries and then calculating the difference, in terms of days, between that value and the date when they played their first match. This way, we have obtained that the average date of the changepoint, for all the 17 countries of interest, falls 14.97 days [95% CI 12.29 to 17.47] after the beginning of their participation in the tournament (approximately two weeks).

Table 5.4: Countries with a changepoint coincidental with a reversal from a decrease to an increase in the SARS-COV-2 case rate that occurred during the European football championship.

| **Country** | $\tau$ | **Diff** | $b_1$ | $b_2$ | $a_1$ |
|---|---|---|---|---|---|
| Austria | 36.4 (35.8, 37.1) | 20 | -0.05 (-0.06, -0.05) | 0.08 (0.08, 0.09) | 6.28 (6.25, 6.32) |
| Belgium | 24.9 (24.6, 25.2) | 10 | -0.06 (-0.06, -0.06) | 0.04 (0.04, 0.04) | 7.70 (7.68, 7.71) |
| Croatia | 28.7 (27.4, 30.3) | 13 | -0.06 (-0.06, -0.06) | 0.02 (0.02, 0.03) | 5.87 (5.83, 5.92) |
| Czechia | 26.2 (25.3, 27.2) | 9 | -0.06 (-0.06, -0.05) | 0.02 (0.02, 0.03) | 6.30 (6.26, 6.34) |
| Denmark | 28.2 (27.8, 28.6) | 13 | -0.06 (-0.06, -0.06) | 0.05 (0.05, 0.06) | 7.13 (7.11, 7.15) |
| Finland | 19.8 (18.4, 21.0) | 5 | -0.03 (-0.04, -0.03) | 0.04 (0.04, 0.05) | 4.98 (4.90, 5.05) |
| France | 34.7 (34.6, 34.8) | 17 | -0.06 (-0.06, -0.06) | 0.11 (0.11, 0.11) | 9.28 (9.28, 9.29) |
| Germany | 35.1 (34.6, 35.5) | 17 | -0.07 (-0.07, -0.07) | 0.05 (0.05, 0.05) | 8.60 (8.59, 8.62) |
| Hungary | 40.3 (38.4, 42.0) | 22 | -0.06 (-0.06, -0.06) | 0.03 (0.02, 0.05) | 5.99 (5.95, 6.03) |
| Italy | 36.5 (36.3, 36.8) | 23 | -0.05 (-0.05, -0.05) | 0.09 (0.09, 0.10) | 8.25 (8.24, 8.26) |
| Netherlands | 26.4 (26.2, 26.6) | 10 | -0.06 (-0.06, -0.06) | 0.09 (0.09, 0.09) | 8.18 (8.17, 8.20) |
| Macedonia | 34.8 (31.9, 37.6) | 19 | -0.05 (-0.05, -0.04) | 0.05 (0.04, 0.07) | 3.59 (3.46, 3.72) |
| Poland | 35.2 (33.5, 36.8) | 18 | -0.07(-0.07, -0.07) | 0.01 (-0.00, 0.01) | 6.92 (6.89, 6.94) |
| Slovakia | 39.4 (36.8, 42.1) | 22 | -0.05 (-0.05, -0.04) | 0.02 (0.00, 0.03) | 5.02 (4.96, 5.08) |
| Spain | 24.9 (24.8, 25.0) | 8 | -0.01 (-0.01, -0.01) | 0.07 (0.06, 0.07) | 8.45 (8.44, 8.46) |
| Switzerland | 32.9 (32.5, 33.5) | 18 | -0.07 (-0.07, -0.07) | 0.08 (0.08, 0.08) | 6.93 (6.91, 6.96) |
| Ukraine | 25.8 (25.1, 26.5) | 10 | -0.05 (-0.05, -0.05) | 0.00 (0.00, 0.01) | 8.06 (8.04, 8.07) |

Now, we have made a step further and, taking the mean values for the coefficients $b_1$ and $b_2$, we have estimated how the slopes for the two lines have changed, on average, before and after the changepoint. We have obtained that all the 17 countries have had a decreasing number of daily cases until the changepoint and ended up with a reversed trend afterwards.

Table 5.5 shows the halving time before and the doubling time after the changepoint, for each given country of this group. More precisely, the mean halving time before the changepoint is 18.07 days [95% CI 11.81 to 29.42], while the mean doubling time after the changepoint is 29.10 days [95% CI 14.12 to 49.78]. The credible intervals are quite wide but if we better investigate the values reported in Table 5.5, we recognize that most of the deviation depends on just three countries, namely: Spain, Ukraine, and Poland, with their exceptionally large values.

To better highlight and summarize all the results we have discussed so far, we also present Figure 5.7 and Figure 5.8, where the same results are portrayed from a clear graphical viewpoint. In particular, Figure 5.7 takes into account the inversion of the SARS-COV-2 case trend of the following countries: Austria, Belgium, Croatia, Czechia, Denmark, Finland. France, Germany, Hungary, Italy. Figure 5.8, instead, shows the inversion of the SARS-COV-2 case trend of Netherlands, North Macedonia, Poland, Slovakia, Spain, Switzerland, and Ukraine.

Table 5.5: Quantifying the inversion from a decrease to an increase in COVID-19 case rate for the countries of Table 5.7

| Country | Halving time $H_b$ | Doubling Time $D_a$ |
|---|---|---|
| Austria | 12.69 | 8.18 |
| Belgium | 11.05 | 17.60 |
| Croatia | 11.77 | 28.32 |
| Czechia | 12.05 | 28.22 |
| Denmark | 11.49 | 12.71 |
| Finland | 21.81 | 15.84 |
| France | 11.86 | 6.32 |
| Germany | 10.14 | 13.17 |
| Hungary | 11.10 | 22.10 |
| Italy | 13.56 | 7.43 |
| Netherlands | 12.11 | 7.69 |
| N. Maced. | 14.88 | 13.20 |
| Poland | 9.67 | 92.80 |
| Slovakia | 14.91 | 41.59 |
| Spain | 103.50 | 10.62 |
| Switzerland | 10.03 | 8.56 |
| Ukraine | 14.43 | 159.00 |

All the figures in this section should be interpreted as follows:

- Yellow space: duration of the tournament. Red vertical line: first match.
- Purple vertical line: last match.
- Green vertical line: last hosted match. Blue vertical line: changepoint.
- Blue space: CI amplitude for the changepoint. Blue bell-shaped peaks: peaks of the probability density function for the changepoint.
- Green segment: case rate trend before the changepoint.
- Red segment: case rate trend after the changepoint.
- Grey segments: fitted lines drawn randomly from the posterior distribution, based on the corresponding CI.
- Black dots: number of daily SARS-COV-2 cases.
- Rightmost $y$-axis: number of cases.
- Leftmost $y$-axis: logarithm of the number of cases

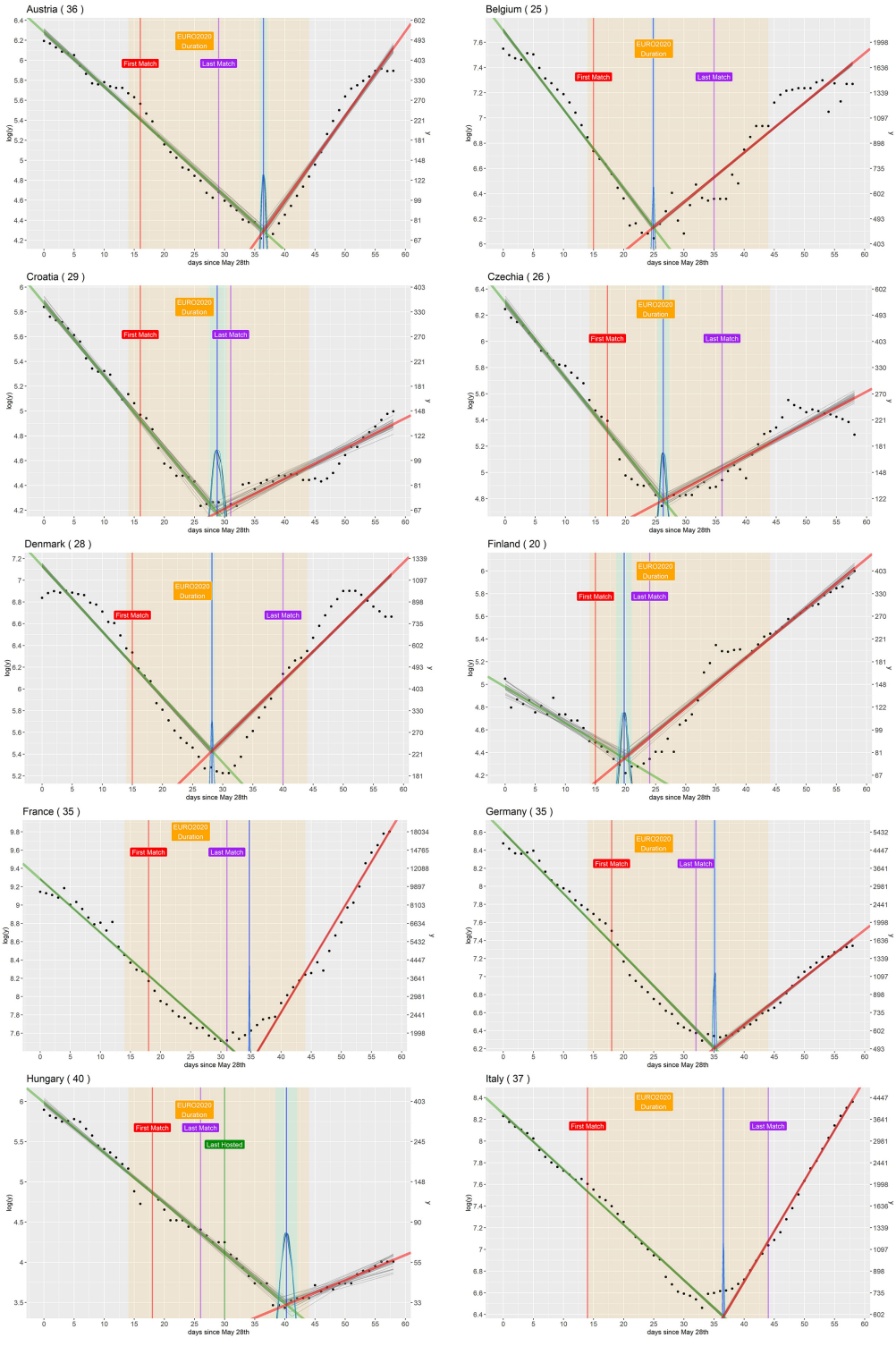



Figure 5.7: Inversion of the SARS-COV-2 case trend for Austria, Belgium, Croatia, Czechia, Denmark, Finland. France, Germany, Hungary, Italy, occurring not later than two to three weeks after their first match.

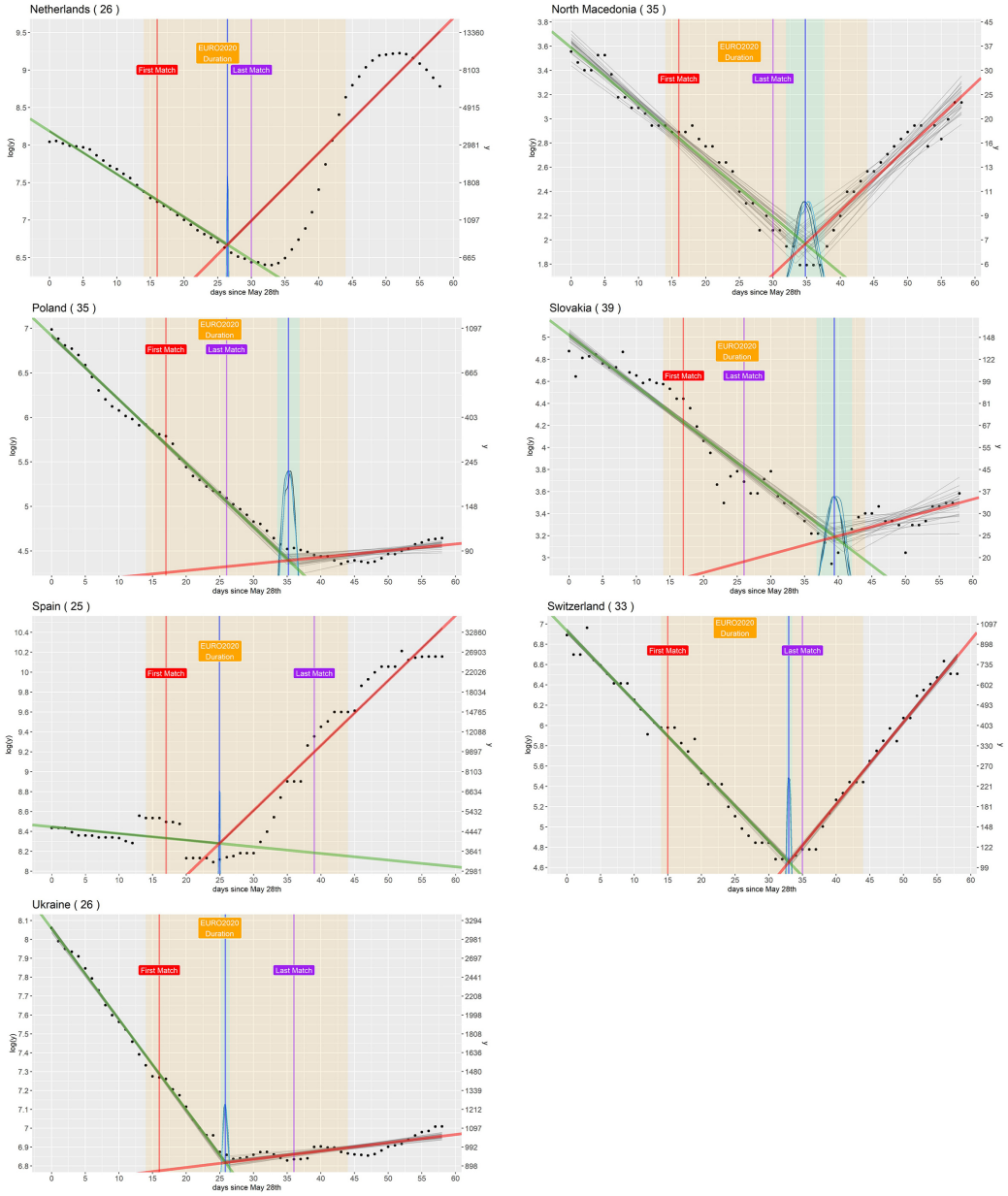



Figure 5.8: Inversion of the SARS-COV-2 case trend for Netherlands, North Macedonia, Poland, Slovakia, Spain, Switzerland, Ukraine, occurring not later than two to three weeks after their first match.

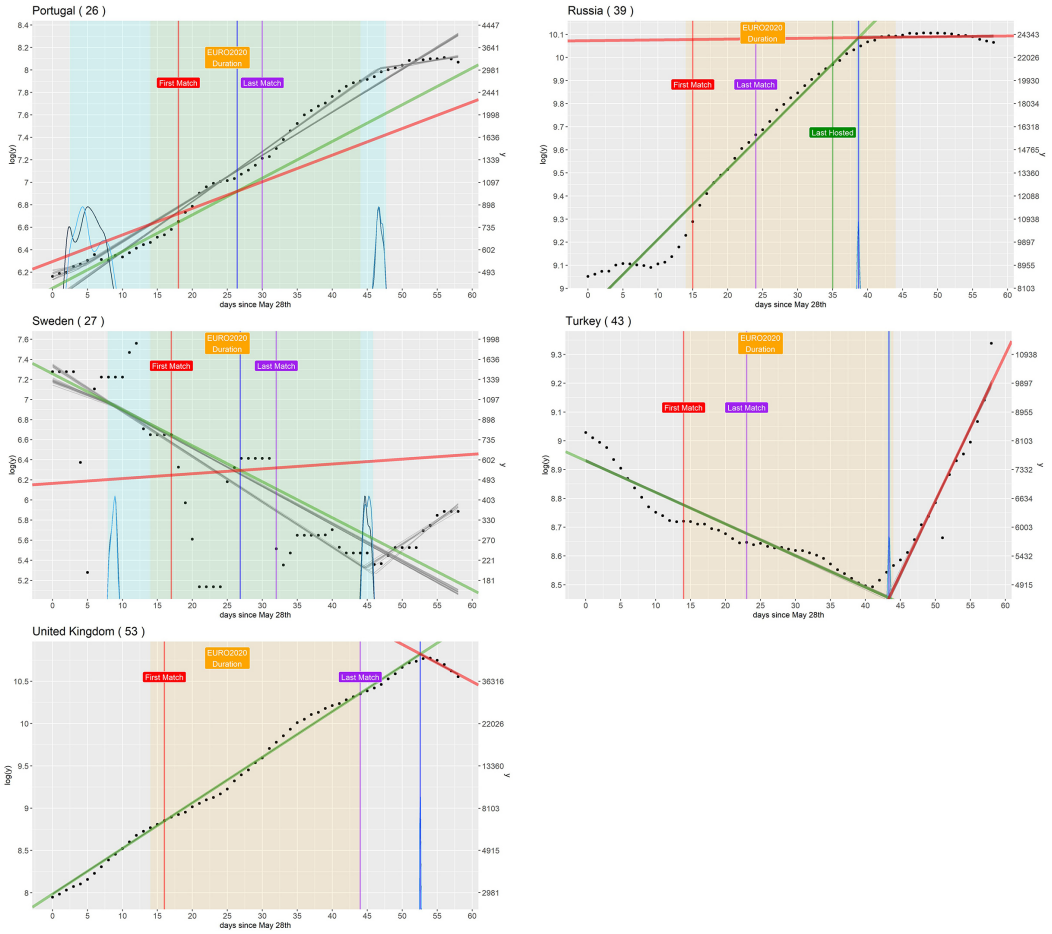



Figure 5.9: Portugal, Russia, Sweden, Turkey, and UK break the pattern, without a well recognizable changepoint or a reversal in the case rate, occurring not later than two to three weeks after the beginning of the tournament.

All the 5 remaining countries (i.e., Portugal, Russia, Sweden, Turkey, and UK), instead, break the pattern and cannot be considered evidence in favor of the research hypothesis. In particular:

(i) Portugal, Russia and UK show a robust increasing trend in the SARS-COV-2 infection case starting well before the beginning the tournament, hence the detected changepoints, as well as the relative slopes, cannot considered an evidence in favor the hypothesis

(ii) Turkey seem to show quite a regular pattern, with a well identifiable changepoint and the usual inverting trend in the case rate, nonetheless the problem here is that that changepoint happens well after the team left the competition, more than four weeks since its first match;

(iii) finally, for Sweden, the model fails to fit because there seems to have two different changepoints, that are either before or after the tournament, making them irrelevant.

The situations mentioned above are illustrated in Figure 5.9, where it is evident that all those five countries break the pattern.

Finally, Table 5.6 reports the value of $\tau$, diff and of all the other parameters, with the corresponding 95% CI. Of particular interest, here, is the large excursion in the CIs for Sweden and Portugal that witness the peculiarity of that situation.

Table 5.6: parameters for the five countries that break the pattern.

| Country | $\tau$ | Diff | $b_1$ | $b_2$ | $a_1$ |
|---|---|---|---|---|---|
| Portugal | 26.4 (2.4, 47.7) | 8 | 0.03 (0.01, 0.05) | 0.08 (0.08, 0.09) | 6.28 (6.25, 6.32) |
| Russia | 38.7 (38.4, 39.0) | 24 | 0.03 (-0.03, -0.03) | 0.00 (0.00, 0.00) | 8.91 (8.90, 8.91) |
| Sweden | 26.9 (7.9, 45.8) | 10 | -0.04 (-0.05, -0.02) | 0.00 (-0.04, 0.05) | 7.26 (7.15, 7.36) |
| Turkey | 43.4 (43.1, 43.6) | 29 | -0.01 (-0.01, -0.01) | 0.05 (0.05, 0.06) | 8.93 (8.92, 8.94) |
| UK | 52.6 (27.8, 28.6) | 37 | 0.05 (0.05, -0.05) | -0.04 (-0.05, -0.04) | 7.99 (7.98, 7.99) |

### Countries That Did Not Participate in the Tournament

While maintaining the pure observational nature of the inferences of our analysis, about the effect of the tournament, we took advantage of another natural experiment, by observing what has happened, during the tournament, in some 12 additional European countries that did not take part in the European football championship (considering the beginning of the tournament as the basis of our statistical observations).

This group was comprised of the following countries (with motivations for their choice reported in brackets): Greece and Ireland (great football traditions),

Table 5.7: Estimated parameters for countries not participating in EURO2020.

| Country | $\tau$ | Diff | $b_1$ | $b_2$ | $a_1$ |
|---|---|---|---|---|---|
| Moldova | 9.6 (6.5, 12.5) | -4 | -0.06 (-0.10, -0.03) | 0.01 (0.01, 0.02) | 4.36 (4.21, 4.50) |
| Norway | 16.6 (15.1, 18.0) | 3 | -0.05 (-0.06, -0.04) | -0.00 (-0.00, 0.00) | 6.03 (5.98, 6.07) |
| Azerbaijan | 24.2 (23.0, 25.3) | 10 | -0.08 (-0.08, -0.07) | 0.06 (0.05, 0.06) | 5.47 (5.41, 5.54) |
| Greece | 26.0 (25.8, 26.3) | 12 | -0.06 (-0.06, -0.06) | 0.07 (0.07, 0.07) | 7.53 (7.51, 7.55) |
| Ireland | 31.6 (30.5, 32.7) | 18 | -0.01 (-0.01, -0.01) | 0.06 (0.05, 0.06) | 6.05 (6.01, 6.08) |
| Serbia | 34.8 (33.5, 36.2) | 21 | -0.05 (-0.05, -0.04) | 0.05 (0.04, 0.06) | 5.83 (5.79, 5.87) |
| Lithuania | 39.1 (38.2, 40.1) | 25 | -0.08 (-0.08, -0.08) | 0.09 (0.08, 0.10) | 6.42 (6.39, 6.45) |
| Latvia | 45.1 (41.5, 48.4) | 31 | -0.05 (-0.05, -0.05) | 0.02 (-0.01, 0.05) | 5.91 (5.87, 5.94) |
| Romania | 37.6 (35.2, 39.9) | 24 | -0.06 (-0.06, -0.06) | 0.04 (0.03, 0.05) | 5.77 (5.72, 5.82) |
| Bosnia | 38.6 (36.2, 40.8) | 25 | -0.06 (-0.06, -0.05) | 0.04 (0.02, 0.06) | 4.63 (4.55, 4.70) |
| Bulgaria | 39.8 (33.0, 44.2) | 26 | -0.04(-0.04, -0.03) | 0.04 (0.01, 0.06) | 5.49 (5.44, 5.55) |
| Iceland | 46.8 (44.9, 48.4) | 33 | -0.01 (-0.02, 0.00) | 0.31 (0.27, 0.36) | 1.40 (1.11, 1.70) |

Romania and Azerbaijan (hosting countries), Norway and Iceland (representatives of Northern Europe), Bulgaria and Moldova (representatives of Eastern Europe), Serbia and Bosnia (representatives of Balkans), Latvia and Lithuania (largest countries representatives of Baltic Europe). The results of the application of our method to the above 12 countries are presented in Table 5.7. The 12 countries are listed based on the increasing value of *diff* (the number of days that separate the changepoint from the beginning of the tournament).

Needless to say, many other countries were left out. The motivations were manifold, ranging from their limited geographical dimensions (e.g., Malta, Faroe Islands, San Marino, Cyprus, Andorra, Montenegro, Kosovo, etc.) to geopolitical considerations, also in relationship with the game of football. For example: Georgia, Armenia, Kazakhstan, and Belarus are not famous for their international football traditions. Moreover, they are also well aligned with the contagion dynamics of one of their most influential neighboring countries, that is, Russia, we had already examined.

Here, it is important to remind what was already stated at the end of Section 5.4.1: COVID-19 cases can show their biggest jumps two to three weeks after a particular mass event, hence only those countries with inverting changes occurring in the time interval from 5 to 23 days after the beginning of the tournament were considered as those that have followed the pattern. This group is comprised of Greece, Azerbaijan, Ireland, Serbia. Just 4 countries out of 12 (33%).

For all the other 8 countries (67%), either their changepoint was premature (Norway, Moldova) or it came too late, precisely more than 23 days after the beginning of the tournament (Latvia, Lithuania, Romania, Bosnia, Bulgaria, and Iceland, in some cases, even without a clear case trend inversion, e.g., Bosnia). As

before, with Figures 5.10 and 5.11 we have portrayed a graphical representation of the same data of Table 5.7 for all the 12 countries of interest.

In conclusion, these final numbers have clearly shown that, while one could suppose that an increase in COVID-19 cases may have been an inevitable consequence of the general European situation in July 2021, the European football tournament, with its mass gatherings, has played the important role of accelerator of this phenomenon, for many of its participating countries.

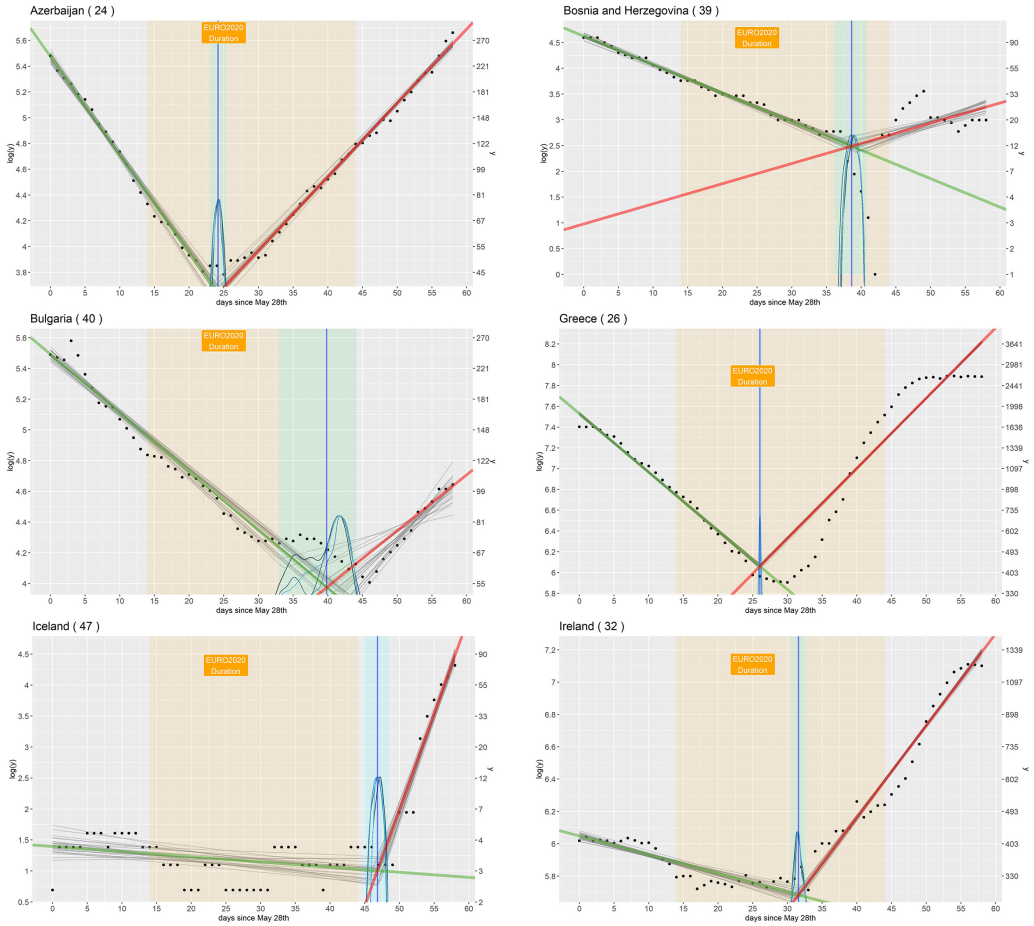



Figure 5.10: Azerbaijan, Bosnia, Bulgaria, Greece, Iceland, Ireland.

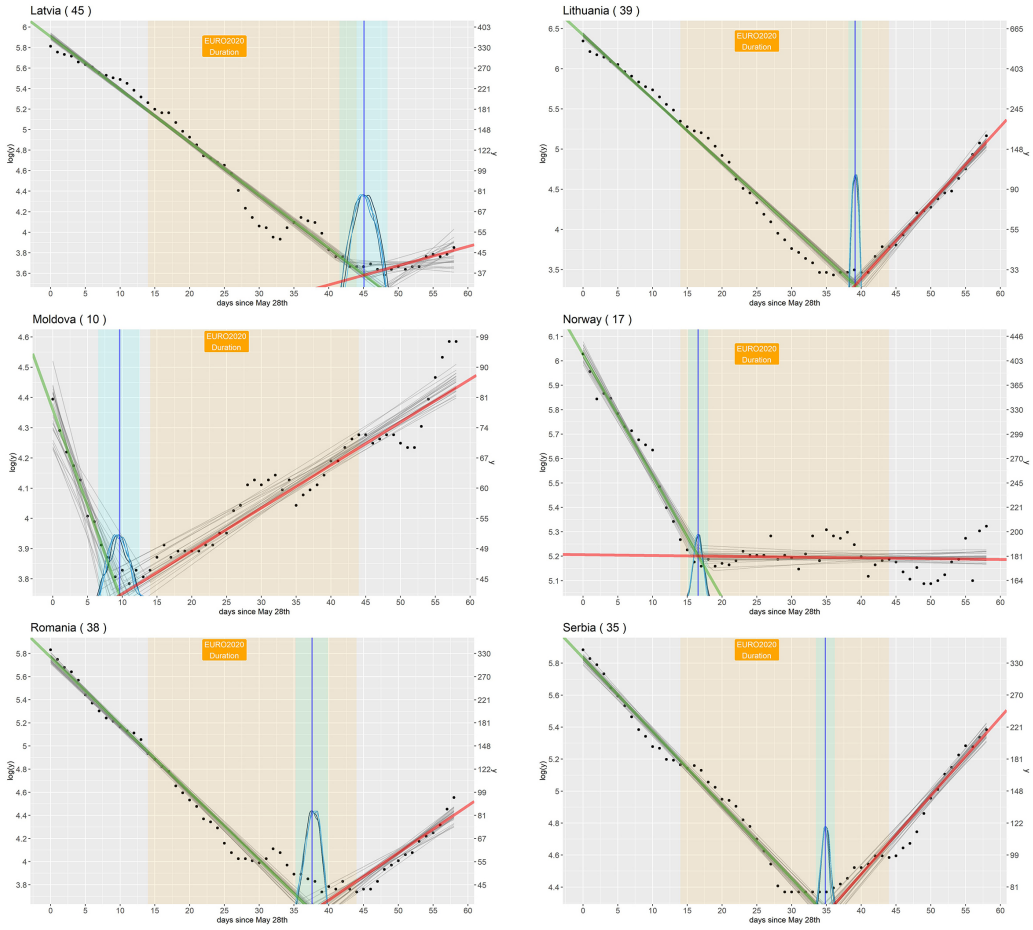



Figure 5.11: Latvia, Lithuania, Moldova, Norway, Romania, Serbia.

## 5.5 Covid Seasonality

At the end of 2021, as the virus had been around for almost two years, a discussion started about the possibility that COVID-19 might follow a seasonal pattern, similar to many other viral infections, like measles and flu, for example. This idea gained momentum probably because of how the contagion receded during the summer months in many Western countries, only to start climbing back up again with the start of autumn, finally reaching a new peak during the winter holidays.

From a scientific perspective, the debate on an infection pattern that repeats over a one-year period has been driven by several analyses investigating the correlation between SARS-COV-2 and various climatic (and environmental) factors, such as temperature, humidity, and UV radiations. The rationale behind this research is that if a negative correlation between SARS-COV-2 and higher temperatures and exposure levels to UV radiation can be demonstrated, then lower COVID-19 infection rates should happen in some given seasons of the year.

D'Amico et al. used a multivariate regression to assess the influence of temperatures and vaccinations on mortality rates in temperate climate countries. They found a negative correlation with temperatures and discovered that the vaccination's effect grew larger as the temperature decreased [209]. Similarly, Ma et al. studied the problem in the United States using a generalized additive mixed model. Their findings are that temperatures are negatively correlated with COVID-19, in an almost linear way, in the range of 20-40° C [210].

However, some other research begins to point out the weaknesses of this type of studies. For example, Fontal et al., while studying the negative correlation between the virus and both higher temperature and humidity, found that there are moments in which this correlation can be inverted, typically corresponding to summer outbreaks in certain regions [211]. The authors suggest that this can be due when various human activities take over, like intensive use of air conditioning, lack of preventive measures and uncontrolled mass gatherings. Also, Sera et al. have expressed their concerns, concluding that the effect of weather, while present, is negligible when compared to the decisive impact of control interventions [212]. Baker et al. have argued that climate factors can play an important factor in the infection when the virus is in the endemic stage. In contrast, during the pandemic stage, it only drives modest changes [213]. Finally, Telles et al, where it has been demonstrated that a combination of factors, including climate, control policies and the use of urban spaces could influence the seasonality of COVID-19 [214].

The positive effect of good weather seems to contrast with several COVID-19 contagions that have broken out, with broad impact, even if with unfavorable climates to the spread of the virus. For example, while this is completely anecdotal,

one could wonder what mechanisms were behind the resurgence of the contagion in May 2020 in Israel [215]. Similarly, the 2021 Olympic Games in Japan took place during the summer when the weather was optimal, but the virus spread, even in the presence of high security standards [216]. In the same period, the European Football Championship took place, and this tournament was connected with an increase of new cases in many involved countries [138]. Also, a new peak hit strong just a few weeks after the president of USA gave a speech during the 4th of July celebration, declaring the final success in beating the pandemic, but a new peak [217]. Finally, the third wave across Europe started at the end of 2021 summer in many eastern countries, when the temperatures were still relatively high.

Following this scientific debate, we investigated the one-year seasonality hypothesis employing a technique from signal processing. Applying a Fourier transform to the daily SARS-CoV-2 infections timeseries at a worldwide level, we looked for peaks in the frequency spectrum that could inform us about the presence, or lack thereof, of cycles in the spread of the virus

### 5.5.1 Methods

The method we adopted for our investigation was a Fourier spectral analysis. In particular, we applied a Discrete Fourier Transform (DFT) to the time series of the number of the new daily SARS-CoV-2 cases and looked for outstanding peaks in the frequency domain corresponding to specific periods [218]. This Fourier frequency spectrum analysis was performed with the precise intent to obtain a converted peak spectrum, indicating the strength and the recurrence of the pandemic waves. In particular, we looked for peaks in the frequency spectrum that could reasonably indicate a periodicity with a certain length. Employing a spectral analysis on the time series of the COVID-19 cases has allowed us to understand, with less ambiguity, the period length of the recurrent outbreaks, instead of counting and observing the number of infections, directly

The 1-dimensional DFT $y[k]$ of length N, of the length-N sequence $x[n]$, is defined as:

$$y[k] = \sum_{n=0}^{N-1} x[n]\, e^{-2\pi j \frac{kn}{N}} \tag{5.9}$$

Where $y[k]$ corresponds to magnitude of the k-th frequency and n represents the n-th day of the time series, with x being the daily number of SARS-COV-2 cases registered on that n-th day of the series.

The period of observation for this study goes from 1 February 2020 to 24 December 24 2021, with the decision not to take into consideration the strong

SARS-CoV-2 outbreak that hit Europe in December 2021, as the progression of this wave was still ongoing in many of the investigated countries during our analysis. The length of the input was set to 730 (two years). Since our study's real period of observation started on 1st February 2020 (until 24th December 2021), the string x was left padded with zeros to reach 730 samples. This zero padding did not alter the validity of the operation since in all the considered countries no SARS-COV-2 infection was registered before 1st February 2020.

To conclude, using a Python library called SciPy [219] we performed a DFT of the time series of the SARS-COV-2 data of each country, that returned all the peaks in the frequency spectrum at their corresponding frequency which can be inverted to obtain the repetition period.

### 5.5.2 Results

The next five figures show the DFTs obtained for all the countries subject of our study, using two separate plots. For each country, the leftmost plot reports the time series of the new daily SARS-COV-2 infections during the observed period. In all plots on the left, $x[n]$ is the COVID-19 timeseries of interest, where $x$ is the number of daily new infections per each day $n$. In all the plots on the right we find the corresponding frequency spectrum output by the DFT. The red line connects the magnitude associated with the different frequencies k, expressed on a yearly basis (i.e., one, twice, trice a year and so on).

Two preliminary facts are noteworthy. First, we can observe a peak in the frequency spectrum representing the 7-day cycle associated with the case reporting process, on the rightmost side of all these DFT plots. This was a quite expected fact, since the reporting process causes an oscillation during the week, in almost all the considered countries. Since we have 52 weeks in a year this peak appears around the location corresponding to this number.

Second, we can observe higher magnitudes and one or more peaks on the opposite side of the spectrum for all our DFT plots. This might be an indication of cyclical patterns in the pandemic, with the first point ($k = 1$) corresponding to a one-year cycle that people commonly associate with seasonal illnesses. However a word of caution is necessary: given the short length of the timeseries, with only two years of data at the time of writing, this analysis is limited by the so-called Fourier uncertainty principle [220]. This is equal to the same concept in quantum physics and tells us that there is an inherent trade-off between time and frequency resolution. In our case, this uncertainty could only be overcome with a longer signal. In practice, the energy associated with lower frequency, like a 1-year cycle, are spread out and difficult to evaluate. This makes sense as a pandemic the

reoccurs every year would only appear twice in our timeseries.

In the plots we have highlighted three different sectors colored in orange, green and pink (from right to left). Those sectors display temporal intervals, respectively, equal to 3-6 months (orange), 6-9 months (green), and 9-12 months (pink). They should be interpreted as follows: If one observes for a certain country the presence of a peak in a given colored sector of the plot (say the green one, for example), this means that country has been hit by a COVID-19 outbreak, which has recurred with a period of 6-9 months. More precisely, if that peak lies on the x axis in correspondence of a value of $k = 2$, this implies that we have had two outbreaks per year in that country.

Coming now to our results, our 30 DFT plots of Figures 55 to 58 reveal that, in the observed period, all the 30 investigated countries have seen the recurrence of at least one COVID-19 wave, repeating over a variable period in the range 3 - 9 months, with a peak of magnitude (roughly equivalent to the number of new infections) at least half as high as that of the highest peak ever experienced since the beginning of the pandemic until December 2021. These findings suggest that strong COVID-19 outbreaks may repeat with cycles of different lengths, without a precisely predictable seasonality of one year.

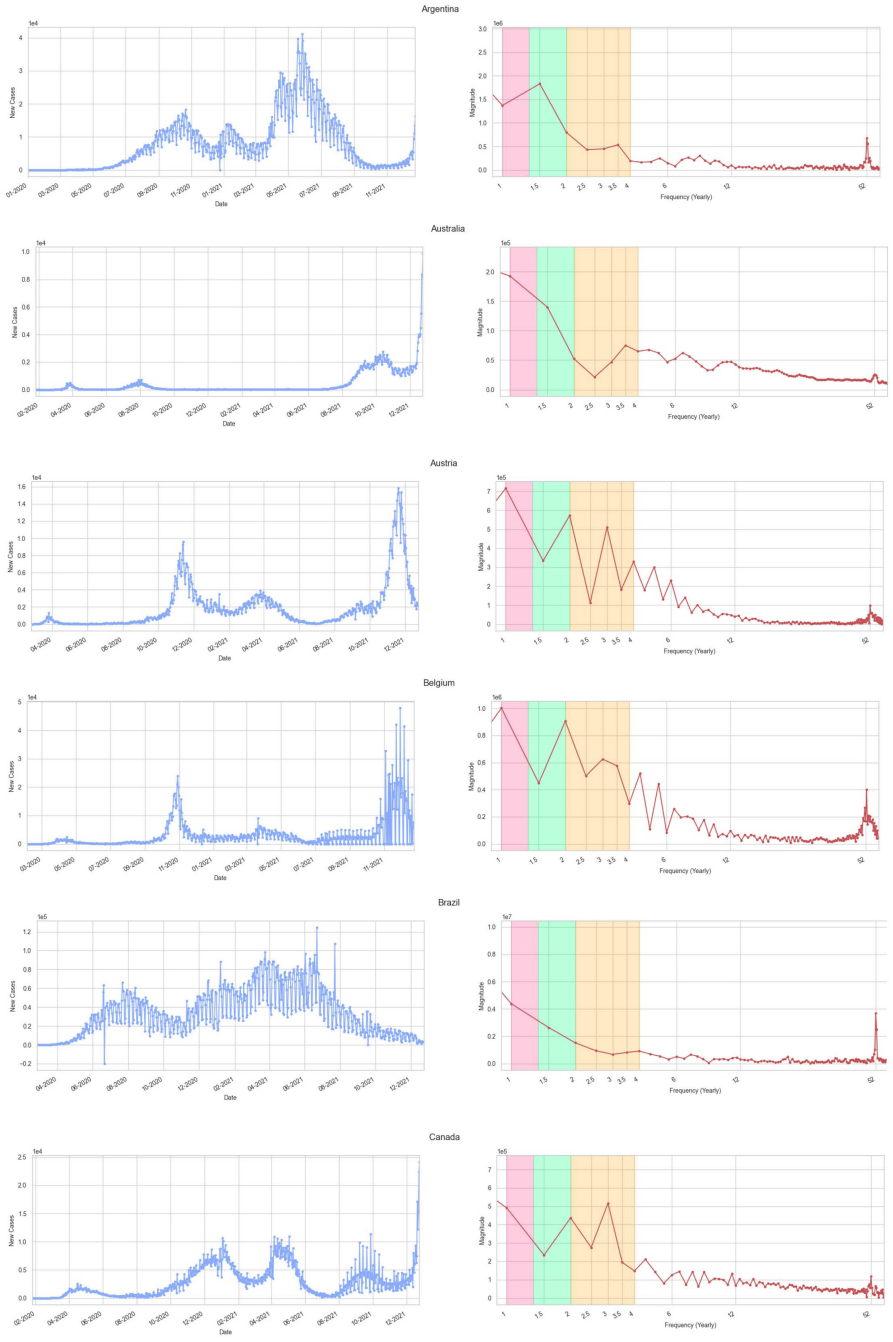



Figure 5.12: DFT plots for Argentina, Australia, Austria, Belgium, Brazil and Canada.

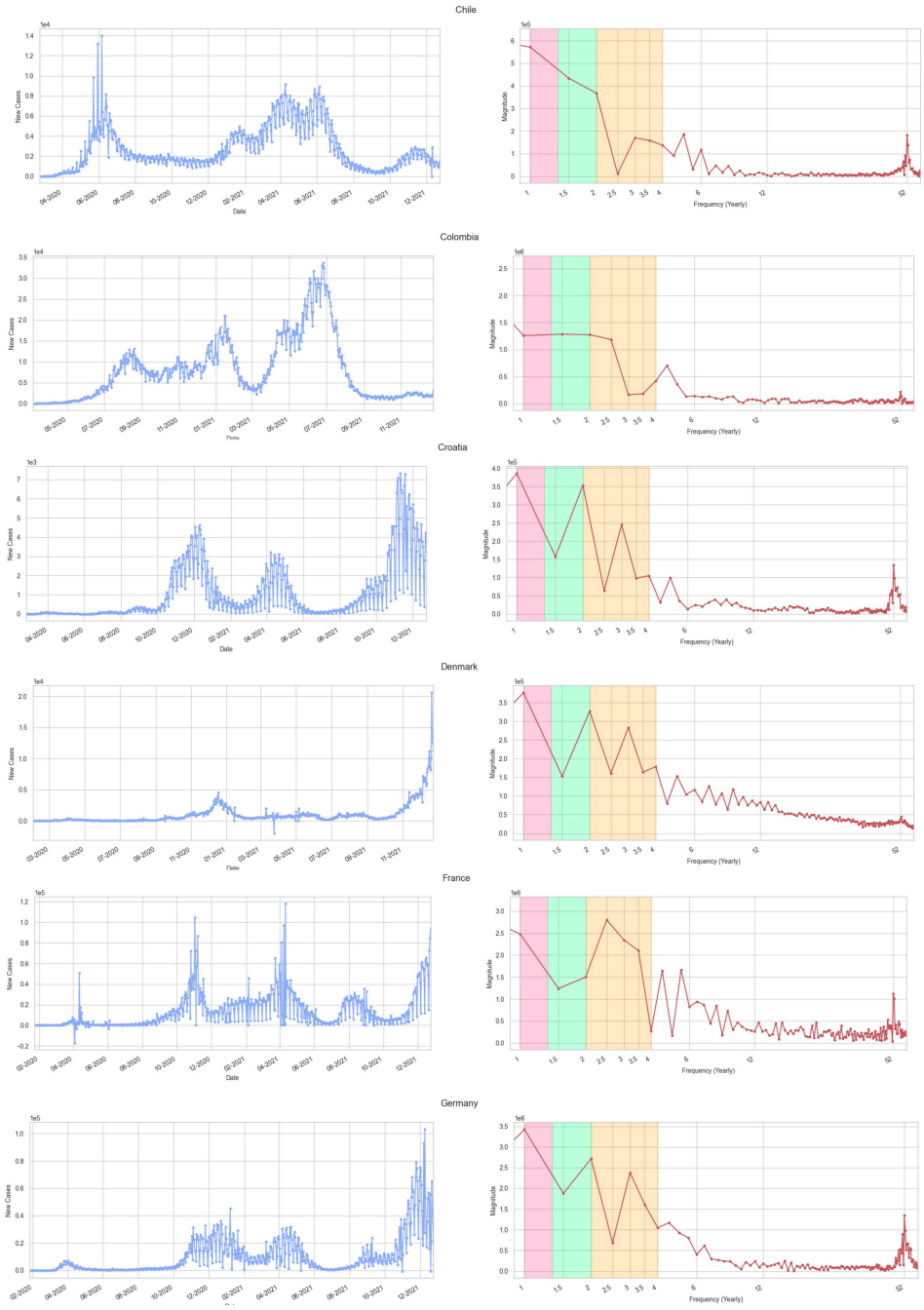



Figure 5.13: DFT plots for Chile, Colombia, Croatia, Denmark, France and Germany

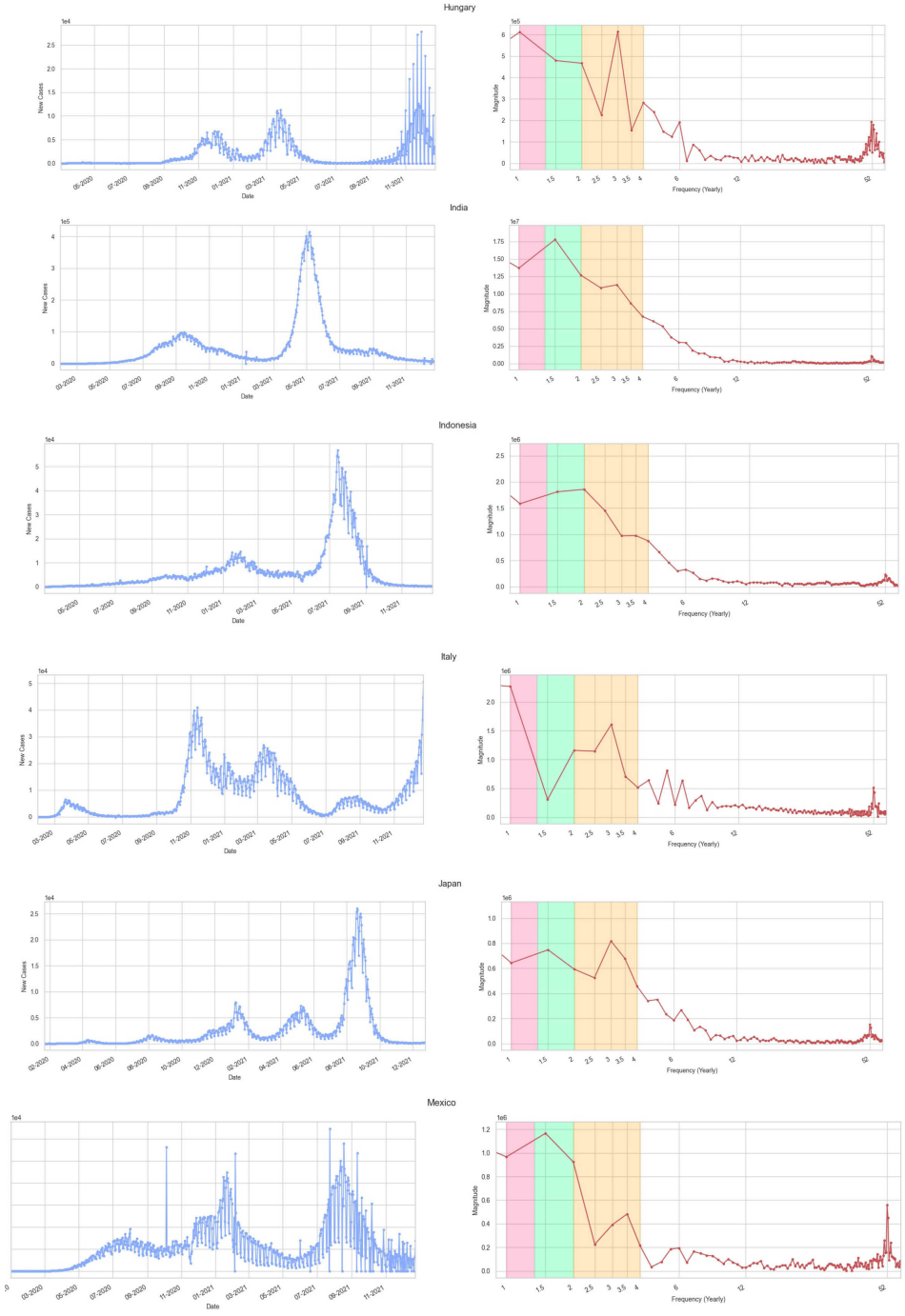



Figure 5.14: DFT plots for Hungary, India, Indonesia, Italy, Japan and Mexico

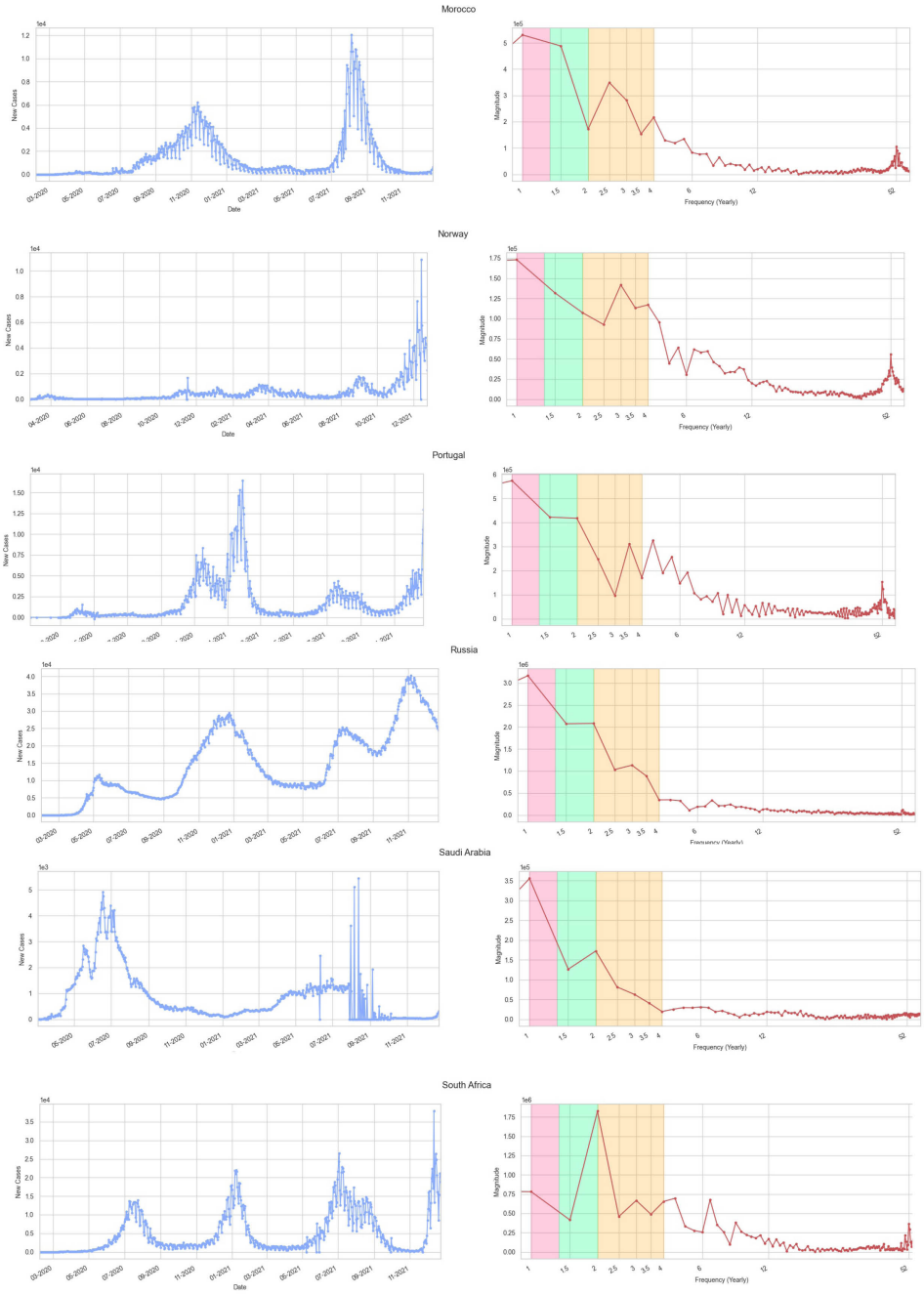



Figure 5.15: DFT plots for Morocco, Norway, Portugal, Russia, Saudi Arabia and South Africa

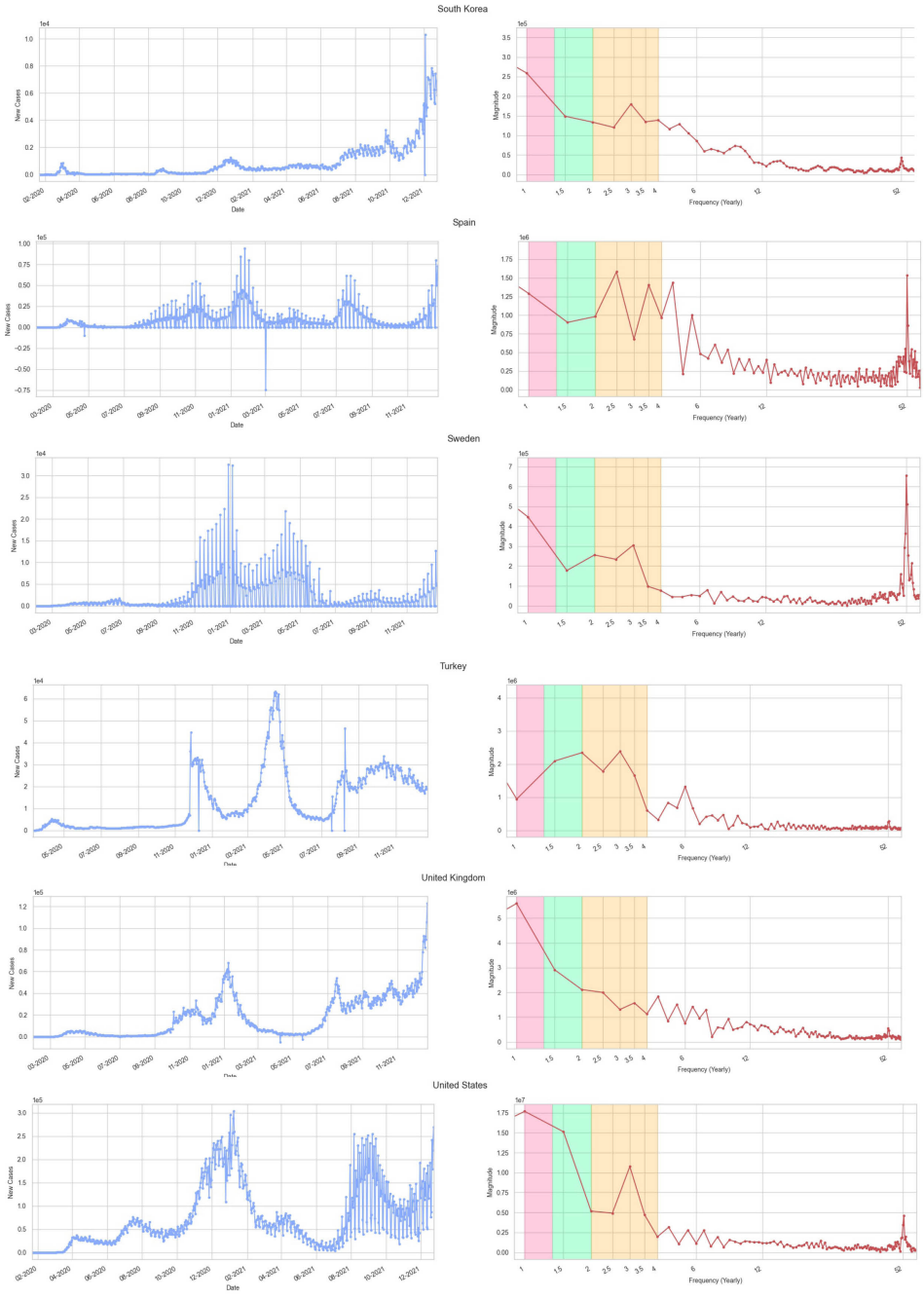



Figure 5.16: DFT plots for for South Korea, Spain, Sweden, Turkey, UK and USA

As an additional analysis suggested in the peer review for this work, we returned to the timeseries of new COVID-19 cases, looking for peaks recurring in each country but adopting a simpler technique. Specifically, using a 7-day rolling average as the raw data of leftmost plots of Figures 5.12 to 5.16 present a weekly periodicity, due to the way COVID-19 tests are carried out and registered, we considered a peak has happened in a given day n, if the number of SARS-COV-2 infections registered in that day was larger than the number of daily SARS-COV-2 cases reported in the 28 days both before and after n. Not only, but to be considered a peak, the number of infections registered on that day n had to be larger than a given threshold computed as the 85% of the average of the daily cases reported in all the days since the beginning of the pandemic until n. Choosing 28 days comes from the working definition of wave as provided in [221], where the three quarters of the upward periods of many studied COVID-19 waves lasted less than a month. Similarly, for the downward periods. The rationale behind the concept of having a threshold came, instead, from the need to filter out all the micro peaks. Finally, upon computation of all the peaks for each country during the period of interest, we chose the two highest ones. Then we computed for each such pair the distance in days between them. Table 5.8 reports the corresponding results.

Precisely, the number of peaks, the distance in days between the two highest ones and their corresponding dates are given for each country. It is interesting to notice that if we average, all over the 30 countries, the values of the temporal distances between the two highest peaks, we obtain a mean of 190 days (SD 100). In other words, we have obtained a confirmation for all our 30 countries of the recurrence of peaks, with an average period of almost 6 months and a standard deviation of nearly 3 months. Moreover, the 80% of the examined countries has that (maximal) temporal distance which falls below the value of one year. Even if we restrict this analysis to only the 13 European countries of our dataset: Austria, Belgium, Croatia, Denmark, France, Germany, Hungary, Italy, Norway, Portugal, Spain, Sweden, and the UK, we achieve an average of almost 5 peaks in a two-years period, with a mean distance between the two highest ones equal to 171 days (SD 85), once again confirming the hypothesis that strong COVID-19 waves may repeat with cycles whose duration break the seasonality pattern of one year.

Table 5.8 also reports, for each registered peak, the variant of the virus that could be considered prevalent at the time of the corresponding outbreak. In particular, to individuate the variant to be associated to each peak, we utilized, for each period and for each country, both the proportion of the total number of sequences collected over time, which fall into some given variant groups, and the corresponding phylogenetic tree. These data are extrapolated, respectively, by the two following initiatives: *Covariants.org* and *Nextstrain.org* [222, 223]. It is worth noticing that both these initiatives are enabled by data shared by the *GI-*

*SAID.org* project that collects all the genome sequences of COVID-19, worldwide [224]. While it is surely of interest the relation between a peak and the frequencies of the sequences collected during a given outbreak, it should be considered that the information about the clades, portrayed in Table 5.8, cannot be always assumed as necessarily representative. The motivation is that genome sampling may not be equal across different countries and periods, with some countries with low sequencing numbers or even with some samples more likely to be sequenced than others. It is worth concluding by pointing out that each mentioned variant in Table 5.8 has been identified based on the conventional names proposed by the genome sequencing initiatives mentioned above (i.e., Covariants.org and Nextstrain.org.). Essentially, each variant's name is comprised of a 2-digit number that represents the year, a progressive alphabetical letter, plus a letter from the Greek alphabet as provided by the WHO organization (e.g., 21J Delta).

Table 5.8: Country, type of climates for that country, number of COVID-19 outbreak peaks, distance in days between the two highest peaks, dates of the two highest peaks, dates of the remaining peaks, and more frequent clades per peak. The mean distance between the two highest peaks is 190 days (SD 100).

| **Country (Climate)** | **highest peaks w/ clades** | **Dist.** | **remaining peaks w/ clades** |
|---|---|---|---|
| Argentina (B/C) | 2020/10/21 (20B/C/D)<br>2021/05/23 (20J Gam., 21G Lamb.) | 214 | 2021/01/11 (20B, 20I Alpha, 20 J Gam.) |
| Australia (A/B/C) | 2020/08/04 (20B/C/F)<br>2021/10/10 (21J Delta) | 432 | 2020/03/30 (19A/B, 20A/B/C) |
| Austria (D/E) | 2020/11/13 (20A)<br>2021/11/24 (21J Delta) | 376 | 2020/03/28 (20A/B/C)<br>2021/04/01 (20I Alpha)<br>2021/09/15 (21J Delta) |
| Belgium (C) | 2020/10/30 (20A/B)<br>2021/03/27 (20I Alpha) | 148 | 2020/04/15 (20A/C)<br>2020/08/12 (20A/C) |
| Brazil (A/C) | 2021/03/27 (20J Gamma)<br>2021/06/22 (20J Gamma) | 87 | 2020/07/28 (20B)<br>2021/01/12 (20B, 20J Gamma) |
| Canada (C/D/E) | 2021/01/09 (20B/C/G)<br>2021/04/12 (20I Alpha, 20J Gam.) | 93 | 2020/04/22 (19A, 20B/C)<br>2021/09/13 (21i/J Delta) |
| Chile (B/C/D) | 2021/04/14 (20J Gam.,21G Lamb.)<br>2021/06/08 (20J Gam., 21G Lamb.) | 55 | 2020/06/12 (19A, 20B/D)<br>2020/10/01 (20B/D)<br>2021/01/25 (20B/D/G, 20I Alpha)<br>2021/11/15 (21J Delta) |
| Colombia (A/C) | 2021/01/20 (19A, 20B/C)<br>2021/06/28 (21H Mu) | 159 | 2020/08/16 (20A/B)<br>2020/11/02 (19A, 20B) |
| Croatia (C) | 2020/12/13 (20B)<br>2021/11/11 (21J Delta) | 333 | 2020/04/01 (20A)<br>2020/07/15 (20B)<br>2020/08/29 (20B)<br>2021/04/21 (20I Alpha) |
| Denmark (D) | 2020/12/18 (20B/E)<br>2021/05/12 (20I Alpha) | 145 | 2020/04/08 (20A/C)<br>2020/09/23 (20A/B/E)<br>2021/03/16 (20I Alpha)<br>2021/08/16 (21J Delta) |
| France (C) | 2020/11/07 (20A)<br>2021/04/17 (20I Alpha) | 161 | 2020/04/18 (19B, 20A)<br>2021/02/11 (20I Alpha)<br>2021/08/16 (21J Delta) |

| **Country (Climate)** | **highest peaks w/ clades** | **Dist.** | **remaining peaks w/ clades** |
|---|---|---|---|
| Germany (C/D) | 2020/12/23 (20A/E)<br>2021/04/25 (20I Alpha) | 123 | 2020/04/02 (19B, 20A/C)<br>2021/09/10 (21J Delta) |
| Hungary (D) | 2020/12/03 (20A)<br>2021/03/26 (20A, 20I Alpha) | 113 | 2020/04/13 (20A) |
| India (A/B/C/D) | 2020/09/16 (20A/B)<br>2021/05/08 (21A/J Delta) | 234 | ... |
| Indonesia (A) | 2021/02/01 (20A/B)<br>2021/07/18 (21I/J Delta) | 167 | 2020/09/26 (20A/B) |
| Italy (B/C) | 2020/11/16 (19A, 20A)<br>2021/03/22 (20E) | 126 | 2020/03/26 (20E, 20I Alpha)<br>2021/01/11 (20I Alpha)<br>2021/08/27 (20J Alpha) |
| Japan (A/C/D) | 2021/01/11 (20B)<br>2021/08/25 (21J Delta) | 226 | 2020/04/15 (19B, 20A/B)<br>2020/08/09 (20B)<br>2021/05/14 (20I Alpha) |
| Mexico (A/B/C) | 2021/01/21 (20A/B/C)<br>2021/08/22 (21I/J Delta) | 213 | 2020/08/01 (20A/B)<br>2020/10/09 (20A/B/C) |
| Morocco (B/C) | 2020/11/17 (20A/B)<br>2021/08/10 (21J Delta) | 266 | 2020/04/22 (20A)<br>2020/06/25 (20A) |
| Norway (D/E) | 2021/03/22 (20I Alpha)<br>2021/09/05 (21I/J Delta) | 167 | 2020/03/29 (19A, 20A/B)<br>2020/11/23 (20A/B/C/E)<br>2021/01/10 (20A/B/E, 20I Alpha)<br>2021/05/26 2021/05/26 (20I Alpha) |
| Portugal(C) | 2020/11/19 (20B/E)<br>2021/01/28 (20E, 20I Alpha) | 70 | 2020/04/03 (20B)<br>2020/07/13 (20B)<br>2021/07/23 (21J Delta) |
| Russia (D/E) | 2020/12/26 (20B/C)<br>2021/11/06 (21J Delta) | 315 | 2020/05/12 (20/B)<br>2021/07/15 (21J Delta) |
| S.Arabia (B) | 2020/06/20 (20A)<br>2021/08/06 (21I Delta) | 412 | 2021/07/02 (21I Delta) |
| S.Africa (B/C) | 2021/01/11 (20A, 20H Beta)<br>2021/07/08 (21I/J Delta) | 178 | 2020/07/19 (20B/D)<br>2021/08/22 (21J Delta) |
| S.Korea (C/D) | 2021/08/15 (21I Delta)<br>2021/09/30 (21I/J Delta) | 46 | 2020/03/04 (19B, 20C)<br>2020/08/27 (20A/C)<br>2020/12/25 (20C)<br>2021/02/20 (20C, 20I Alpha, 21D Eta)<br>2021/04/23 (20A, 20I Alpha) |
| Spain (B/C) | 2021/01/26 (20E, 20I Alpha)<br>2021/07/19 (21J Delta) | 174 | 2020/03/31 (19B, 20A/B)<br>2020/11/04 (20E)<br>2021/04/27 (20I Alpha) |
| Sweden (D/E) | 2021/01/11 (20A/E)<br>2021/04/12 (20I Alpha) | 91 | 2020/04/29 (19A, 20B/C)<br>2020/06/18 (20B) |
| Turkey (B/C/D) | 2020/12/02 (20A/B, 20I Alpha)<br>2021/04/20 (20I Alpha, 20H Beta) | 139 | 2020/04/16 (20A)<br>2021/08/15 (21A/J Delta)<br>2021/10/15 (21J Delta) |
| UK (C) | 2021/01/09 (20I Alpha)<br>2021/07/21 (21J Delta) | 193 | 2020/04/22 (19A, 20D)<br>2020/11/16 (20B/E)<br>2021/09/08 (21J Delta)<br>2021/10/23 (21J Delta) |
| USA (A/B/C/D/E) | 2021/01/11 (20B/C/G, 21C Epsilon)<br>2021/09/13 (21J Delta) | 245 | 2020/04/10 (19A/B, 20B/C)<br>2020/07/22 (20B/C/G)<br>2021/04/14 (20I Alpha) |

## 5.6 Conclusion

This chapter is wholly focused on the topic of the COVID-19 pandemic and, contrary to the others, lacks the presence of deep learning methods, save for some sporadic mentions. However, data science should be about providing value with analysis and extracting information from the data we have available.

As Cynthia Rudin noticed, black-box models can hide all sort of biases and problems inside of them, and in high-risk situations this characteristic is all but desirable [30]. Given the lack of coherent data caused by the discrepancy in collection methodologies we discussed in section 5.1 and, especially in the first months of the pandemic, the lack of data of any kind, we decided to focus on observational studies. That is, a study that organizes and analyzes the available data and that proposes a hypothesis on the processes that lie behind it. Obviously, the conclusion we can draw from such studies are limited. But this limitation stems from the limitation in the observations themselves and we could argue that trying to squeeze more out of what is available can results in misleading conclusions, or in conclusions that are equally flawed and uncertain but perhaps hide the shortcomings behind a wall of mathematical complexity.

All of our studies were not conclusive but suggested further investigation and advised policymakers to take exceptional care given the situation we were going through.

In Section 5.2 we tried using a simple but very easily interpretable regression analysis to have a statistical confirmation of our suspect that tourism was driving a second wave of COVID-19 in Italy. Up until the end of summer 2020, most regions had close to 0 cases and then suddenly there was an increase almost everywhere, and strangely enough also in remote insular regions like Sardinia. Our liner regression was not very good for prediction but was useful to find the most relevant factors. The method showed a strong correlation between tourism data and COVID-19 cases, even when controlling for other factors like aging population and healthcare expenditure. This conclusion was not difficult to draw even without mathematical tools. However, prevention policies had to clash with the pressures from the economy which in the end had the upper hand. That answer to which is the best thing to do is better left to philosophers and political scientist but after more than 2 years we can safely assume that our hypothesis was correct.

Section 5.3 dealt with the matter of school reopening in Italy. The topic spurred endless discussion in every country, and each took a slightly different approach. On our part what we wanted to show is that there were clear clues that suggest a link between school activity and the rise of new cases on a temporal level. We highlighted how going back to school at all levels meant that 11 million Italians

(who are not only children but also adults working as teacher, administration staff, public transportation drivers, etc.) were moving and interacting each day while most workers were still working from home (the increase in people going to the office was 2.5 million). The technique we used is very simple and the results are akin to a visual analysis but still we find it compelling enough to suggest randomized tests in schools. Albeit the Italian government never pushed for such an experiment making it impossible to know the impact of school contagion, also because contact tracing never took off. With the arrival of new variants in following year we saw an ever-increasing number of children testing positive which in a way confirmed the hypothesis of their role in the spread. gain, the decision to not open school clashes with the developmental damages this caused to a whole generation, and we cannot tell which decision is best.

Using the same methodology, in Section 5.4 we also looked at the European Football championship of 2020, which was rescheduled to 2021. This edition was held in different European countries for each match as it was a special edition for the 60th anniversary. We found that, while many countries saw an increase in cases in that period, it is way more consistent in those that participated in the tournament. While the study is quite limited and we it is hard to believe it could sway an organization such as UEFA, we wanted to contribute to the increasing literature on the risk of massive gatherings like sporting events.

Finally, Section 5.5 sought to add a simple yet effective visual confirmation to the fact that the COVID-19 pandemic is not a virus with the traditional yearly seasonality of other coronaviruses like the flu. Applying spectral analysis to the series of new cases we showed how the strongest "frequency" is often the weekly pattern due to data collection. When there are other prevalent components, their frequency often does not correspond to 1 year but rather is shorter. While some epidemiologist suggested that if and when the COVID-19 will become endemic it might have this seasonal pattern for the moment is not a factor to take into account when decided how to act with public policies.

To circle back to the concepts of this thesis, we started with very limited and problematic data and decided to use methods that best fit these limitations. Concentrating on the data meant leaving behind complex machine learning and deep learning models and use simple but effective techniques. This is a data-centric approach to the problem. The results of our methodologies are easily interpretable, and their limitations are clear. The graphical nature of Bayesian changepoint estimation in particular is very easy to understand. In a time of great uncertainty and confusion, even for those who have to decide for the whole population, providing this sort of analysis is useful as it consider the human-factor in approaching a data problem, making it a human-centric result.

# Chapter 6

# Conclusions

Deep learning has quickly moved from being a cutting edge research topic to a tool at the disposal of data scientists employed in companies at every level, all over the world. It has entered virtually any field of application and took over the previous state of the art in most cases. This thesis, as broad as it is, is a demonstration of this incredible diversity.

The widespread use of these technologies is starting to show how big of an impact they can have on our society. This is true, however, both when things work as intended as when they do not. It is very difficult to account for every potential source of ill effects, but shifting the perspective in the design process can go a long way. Human and Data-centric approaches come into play exactly to address this issue.

The success of a system based on statistical methods like deep learning is tied, obviously, to the data it learns from: not only their quantity, but mostly their quality and their meaning. Data should be seen as the programming language of those huge universal approximation machines that are neural networks. The way a dataset is built, together with the choice of loss function, shapes the question the machine should answer. If we are not careful, this question could be biased in ways that may not be evident, especially to the engineers building the system.

To solve this we need the help of human experts, which should always be part of the design and implementation loop of such intelligent systems. Their knowledge can be helpful on many levels: they can help in creating a better dataset, or in fixing and existing one that is imperfect, by highlighting the biases and idiosyncrasies of the data; they can provide insight on how to structure the model and make it better and, most importantly, they can help evaluating the results that the model produces, which often are badly summarized by metrics and loss

functions. When a system is integrated in a human process its value can go beyond the simple accuracy, as there are other "dimensions" where it can be helpful, like saving time or kick-starting a creative endeavor. All this considered, we have shown and discussed four different applications of machine learning and data science.

This final short chapter summarizes the contributions of this thesis and the conclusions which emerged in each part, addressing the research questions we presented in the introduction.

**Section 2** described a model to classify faulty metering device for a water supply company. Addressing **RQ1**, we have seen how distilling the process that generates the data and the experts' knowledge into our dataset and model helped achieving good performance and a model general enough to work on different datasets. Concerning **RQ4**, we showed how visualizing the data distribution helped us define a better way to deal with categorical data. Partially relating to **RQ2** and **RQ3**, in the final section we discussed how the models could be integrated into the established practices of the company. It emerged how their utility goes beyond the AUC score and we need to take into account what the company values.

**Section 3** instead focused on an intelligent system that helped archaeologists find potential sites to survey in the Mesopotamian floodplain. Their process involves manually checking satellite and aerial imagery that covers an enormous area and pinpointing the location of those sites. Fully answering **RQ1**, we saw how their expertise was fundamental on many levels. First they helped us define how to treat the dataset that they annotated and provided, both concerning the satellite images and the sites shapes, and to refine the examples by removing the most problematic instances from the training set. They helped us realize the better approach (segmentation vs. classification) and select hyperparamters. Most importantly they performed human-in-the-loop evaluation that allowed us to better gauge the model performance. This last point ties into **RQ2**, as we soon realized that the usual metrics are limited in this context. Namely, a prediction that points to an unseen site is computed as an error even when its not and a missing prediction can be justified by a site no longer visible in satellite photos. Re-evaluating errors in the test set with an archaeologist showed that the model performance are higher than what initially appeared. Moreover, even when mistakes are present we have to consider their meaning and impact. Concerning **RQ3**, When we asked our colleagues archaeologists, they highlighted how the most important feature of the system is the time that it saves them and also how a lot of what could be considered mistakes are actually something that they consider valuable and would check manually. Finally, we also imagined a new workflow for remote sensing tasks which leverages our model. Beside providing automatic predictions, its use can result in

a continuous refinement of the existing dataset and the production of an overlay, which can guide the eye of domain experts, in a synergy with their knowledge, without replacing them. This visualization provides another aspect to **RQ4**.

**Section 4** documented the results of the work done during a visiting period at KTH Stockholm in the MUSAiC project. The goal was creating a transformer model for traditional folk music and compare it to the previous SotA folkRNN. Answering **RQ1** and **RQ4**, the design and implementation process did not run as smooth as we anticipated and through the use of human expertise (as both musicians and machine learning experts) and visualization techniques we were able to address all the issues and obtain a new state-of-the-art model. Especially important was the use of human evaluation, conducting periodic blind tests against folkRNN to check if the model improved, as the loss function did not provide useful information. This provides yet another answer to RQ3. In the last section we discussed how the model can be used in a setting of human-AI collaboration for music co-creation. In this setting the model provides a new tune that musician can play and refine freely. Concerning **RQ3**, folk music is not as rigid as, for example, classical music and small mistakes in the output can be easily ignored by musicians as long as the musical idea is intelligible.

**Section 5** took a slight detour in dealing with COVID-19 and its relationship with tourism, schools and mass-events as well as its supposed seasonal behavior. These observational studies provided results that were far from conclusive but nonetheless highlighted, in a simple and often graphical way, relationships that were in the data, urging for further experiments or new policies by the government. The approach we took had the goal of contributing to the public and scientific discussion, while keeping in mind the limitations of the data and trying to communicate unambiguous hypothesis without hiding uncertainty. Following **RQ5**, we chose to use only classical methods from statistics that guarantee transparency and interpretability, and can clearly highlight uncertainty in the case of Bayesian techniques. To communicate this efficiently, we focused especially on good visualizations, addressing another aspect of of **RQ4**. Finally, relating to **RQ2**, even if our models were not good in the sense of predictive capabilities, they provided valuable insight and highlighted data and factors that could have helped in further, more conclusive, studies.

# Bibliography


[1] *2018 ACM A.M. Turing Award Laureates*. URL: https://awards.acm.org/about/2018-turing (visited on 10/07/2022).

[2] James Moor. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years". In: *AI Magazine* 27.4 (Dec. 15, 2006). Number: 4, pp. 87–87. ISSN: 2371-9621. DOI: 10.1609/aimag.v27i4.1911. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1911 (visited on 10/06/2022).

[3] Thomas H. Davenport and D. J. Patil. "Data Scientist: The Sexiest Job of the 21st Century". In: *Harvard Business Review* (Oct. 1, 2012). Section: Analytics and data science. ISSN: 0017-8012. URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (visited on 10/07/2022).

[4] Jorge Barreto. "Neural network learning: a new programming paradigm?" In: *Proceedings of the 1990 ACM SIGBDP conference on Trends and directions in expert systems*. 1990, pp. 434–446.

[5] A. Dix et al. *Human-computer Interaction*. Pearson/Prentice-Hall, 2003. ISBN: 978-0-13-046109-4. URL: https://books.google.it/books?id=IuQxui8GHDcC.

[6] Ray Kurzweil. *The singularity is near: When humans transcend biology*. Penguin, 2005.

[7] Elizabeth Svoboda. "Artificial intelligence is improving the detection of lung cancer". In: *Nature* 587.7834 (Nov. 18, 2020). Bandiera_abtest: a Cg_type: Outlook Number: 7834 Publisher: Nature Publishing Group Subject_term: Cancer, Software, Imaging, Machine learning, S20–S22. DOI: 10.1038/d41586-020-03157-9. URL: https://www.nature.com/articles/d41586-020-03157-9 (visited on 10/07/2022).

[8] Sean D. Holcomb et al. "Overview on DeepMind and Its AlphaGo Zero AI". In: *Proceedings of the 2018 International Conference on Big Data and Education*. ICBDE '18. New York, NY, USA: Association for Computing Machinery, Mar. 9, 2018, pp. 67–71. ISBN: 978-1-4503-6358-7. DOI: 10.1145/3206157.3206174. URL: https://doi.org/10.1145/3206157.3206174 (visited on 10/07/2022).

[9] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (Aug. 2021). Number: 7873 Publisher: Nature Publishing Group, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 10/07/2022).

[10] Richard Sutton. "The bitter lesson". In: *Incomplete Ideas (blog)* 13 (2019), p. 12.

[11] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[12] Scott Reed et al. "A generalist agent". In: *arXiv preprint arXiv:2205.06175* (2022).

[13] Stuart J. Russell. *Human compatible: artificial intelligence and the problem of control*. New York?: Viking, 2019. 1 p. ISBN: 978-0-525-55862-0.

[14] Marvin Minsky, Ray Kurzweil, and Steve Mann. "The society of intelligent veillance". In: *2013 IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life*. IEEE, 2013, pp. 13–17.

[15] Eliza Strickland. "Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big". In: *IEEE Spectrum* 59.4 (Apr. 2022). Conference Name: IEEE Spectrum, pp. 22–50. ISSN: 1939-9340. DOI: 10.1109/MSPEC.2022.9754503.

[16] Ben Shneiderman. "Human-centered artificial intelligence: Reliable, safe & trustworthy". In: *International Journal of Human–Computer Interaction* 36.6 (2020). ISBN: 1044-7318 Publisher: Taylor & Francis, pp. 495–504.

[17] Fabio Massimo Zanzotto. "Human-in-the-loop artificial intelligence". In: *Journal of Artificial Intelligence Research* 64 (2019), pp. 243–252.

[18] Xingjiao Wu et al. "A survey of human-in-the-loop for machine learning". In: *Future Generation Computer Systems* 135 (Oct. 1, 2022), pp. 364–381. ISSN: 0167-739X. DOI: 10.1016/j.future.2022.05.014. URL: https://www.sciencedirect.com/science/article/pii/S0167739X22001790 (visited on 10/07/2022).

[19] Andreas Holzinger. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2 (2016). Publisher: Springer, pp. 119–131.

[20] Besmira Nushi et al. "On human intellect and machine failures: Troubleshooting integrative machine learning systems". In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[21] Jianzheng Liu et al. "Rethinking big data: A review on the data quality and usage issues". In: *ISPRS journal of photogrammetry and remote sensing* 115 (2016). Publisher: Elsevier, pp. 134–142.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. event-place: San Francisco, California, USA. New York, NY, USA: ACM, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: http://doi.acm.org/10.1145/2939672.2939778 (visited on 12/03/2019).

[23] Bob L. Sturm. "A simple method to determine if a music information retrieval system is a "horse"". In: *IEEE Transactions on Multimedia* 16.6 (2014). ISBN: 1520-9210 Publisher: IEEE, pp. 1636–1644.

[24] Jiakai Wang. "Adversarial Examples in Physical World." In: *IJCAI*. 2021, pp. 4925–4926.

[25] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. "FIND: Human-in-the-Loop Debugging Deep Text Classifiers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 332–348.

[26] Bob Sturm. "What do these 5,599,881 parameters mean?: an analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer". In: *International Conference on Computational Creativity*. 2018.

[27] Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. "Artificial intelligence and human trust in healthcare: focus on clinicians". In: *Journal of medical Internet research* 22.6 (2020). Publisher: JMIR Publications Inc., Toronto, Canada, e15154.

[28] Elisabetta Lalumera, Stefano Fanti, and Giovanni Boniolo. "Reliability of molecular imaging diagnostics". In: *Synthese* 198.23 (2021). Publisher: Springer, pp. 5701–5717.

[29] Georg Macher et al. "Dependable Integration Concepts for Human-Centric AI-Based Systems". In: *Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops*. Ed. by Ibrahim Habli et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 11–23. ISBN: 978-3-030-83906-2. DOI: 10.1007/978-3-030-83906-2_1.

[30] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019). Publisher: Nature Publishing Group, pp. 206–215.

[31] Marianne Promberger and Jonathan Baron. "Do patients trust computers?" In: *Journal of Behavioral Decision Making* 19.5 (2006), pp. 455–468.

[32] Francisco Maria Calisto et al. "Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification". In: *International Journal of Human-Computer Studies* 150 (2021), p. 102607.

[33] Andrew Dennis Smith et al. "Multi-institutional comparative effectiveness of advanced cancer longitudinal imaging response evaluation methods: Current practice versus artificial intelligence-assisted." In: *Journal of Clinical Oncology* 38.15_suppl (2020), pp. 2010–2010. DOI: 10.1200/JCO.2020.38.15_suppl.2010.

[34] L. Casini et al. "Deep water: Predicting water meter failures through a human-machine intelligence collaboration". In: *Human Interaction and Emerging Technologies*. Vol. 1018. 2020, pp. 688–694. ISBN: 978-3-030-25629-6. DOI: 10.1007/978-3-030-25629-6_107.

[35] M. Roccetti et al. "A Cautionary Tale for Machine Learning Design: why we Still Need Human-Assisted Big Data Analysis". In: *MOBILE NETWORKS AND APPLICATIONS* 2020 (2020), pp. 1–9. DOI: 10.1007/s11036-020-01530-6.

[36] Marco Roccetti et al. "A paradox in ML design: less data for a smarter water metering cognification experience". In: *proceedings of the 5th EAI international conference on smart objects and Technologies for Social Good*. 2019, pp. 201–206.

[37] M. Roccetti et al. "Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures". In: *JOURNAL OF BIG DATA* 6 (2019), pp. 1–23. DOI: 10.1186/s40537-019-0235-y. URL: https://link.springer.com/article/10.1186/s40537-019-0235-y#citeas.

[38] Giovanni Delnevo, Marco Roccetti, and Luca Casini. "Categorical data as a stone guest in a data science project for predicting defective water meters". In: *Proceedings SCIFI-IT'2020 - 4th Annual Science Fiction Prototyping Conference*. Ghent: Eurosis, 2020, pp. 24–26. ISBN: 978-94-92859-10-5. URL: http://arxiv.org/abs/2102.03284v1.

[39] M. Roccetti et al. "Dimensionality Reduction and the Strange Case of Categorical Data for Predicting Defective Water Meter Devices". In: *Human Interaction, Emerging Technologies and Future Applications III*. Vol. 1253. Cham Switzerland: Springer Nature, 2021, pp. 155–159. ISBN: 978-3-030-55306-7. DOI: 10.1007/978-3-030-55307-4_24.

[40] Marco Roccetti et al. "An alternative approach to dimension reduction for pareto distributed data: a case study". In: *JOURNAL OF BIG DATA* 8 (2021), pp. 1–23. DOI: 10.1186/s40537-021-00428-8.

[41] G. Cappiello et al. "Human matters: una applicazione di machine Learning alla fornitura di servizi". In: *Referred Electronic Conference Proceedings of Sinergie - SIMA Management Conference - Management and Sustainability: Creating shared value in the digital era*. Roma: Sinergie, 2019, pp. 313–316. ISBN: 978-88-943937-1-2. DOI: 10.7433/SRECP.EA.2019.55. URL: http:www.sijm.it.

[42] Alison M St. Clair and Sunil Sinha. "State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models!" In: *Urban Water Journal* 9.2 (2012). Publisher: Taylor & Francis, pp. 85–112.

[43] Katarzyna Pietrucha-Urbanik. "Failure prediction in water supply system–current issues". In: *International Conference on Dependability and Complex Systems*. Springer, 2015, pp. 351–358.

[44] Stefano Alvisi et al. "Wireless middleware solutions for smart water metering". In: *Sensors* 19.8 (2019). Publisher: MDPI, p. 1853.

[45] SE Roberts and IR Monks. "Fault detection of non-residential water meters". In: *MODSIM2015, 21st international congress on modelling and simulation. Modelling and simulation society of Australia and New Zealand*. 2015, pp. 2228–33.

[46] Iñigo Monedero et al. "An approach to detection of tampering in water meters". In: *Procedia Computer Science* 60 (2015). Publisher: Elsevier, pp. 413–421.

[47] Valerie Vaquet et al. "Taking Care of Our Drinking Water: Dealing with Sensor Faults in Water Distribution Networks". In: *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 682–693.

[48] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3 (Mar 2003), pp. 1157–1182.

[49] Zhidong Li and Yang Wang. "Domain knowledge in predictive maintenance for water pipe failures". In: *Human and machine learning*. Springer, 2018, pp. 437–457.

[50] François Chollet et al. *Keras*. 2015. URL: `https://github.com/fchollet/keras`.

[51] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: `https://www.tensorflow.org/`.

[52] Kyunghyun Cho et al. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 2014, pp. 103–111.

[53] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. Issue: 1. Citeseer, 2013, p. 3.

[54] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014). Publisher: JMLR. org, pp. 1929–1958.

[55] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[56] Joffrey L Leevy et al. "A survey on addressing high-class imbalance in big data". In: *Journal of Big Data* 5.1 (2018). Publisher: Springer, pp. 1–30.

[57] Alaa Tharwat. "Classification assessment methods". In: *Applied Computing and Informatics* (2020). Publisher: Emerald Publishing Limited.

[58] Yury Gorishniy et al. "Revisiting deep learning models for tabular data". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18932–18943.

[59] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?" In: *arXiv preprint arXiv:2207.08815* (2022).

[60] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[61] Gerard V Trunk. "A problem of dimensionality: A simple example". In: *IEEE Transactions on pattern analysis and machine intelligence* 3 (1979). Publisher: IEEE, pp. 306–307.

[62] Tallha Akram et al. "A multilevel features selection framework for skin lesion classification". In: *Human-centric Computing and Information Sciences* 10.1 (2020). Publisher: Springer, pp. 1–26.

[63] Alex Pappachen James and Sima Dimitrijev. "Ranked selection of nearest discriminating features". In: *Human-Centric Computing and Information Sciences* 2.1 (2012). Publisher: Springer, pp. 1–14.

[64] Yanning Shen, Morteza Mardani, and Georgios B Giannakis. "Online categorical subspace learning for sketching big data with misses". In: *IEEE Transactions on Signal Processing* 65.15 (2017). Publisher: IEEE, pp. 4004–4018.

[65] Hirotaka Niitsuma and Takashi Okada. "Covariance and PCA for categorical variables". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2005, pp. 523–528.

[66] Michael J Greenacre. "Correspondence analysis". In: *The Oxford Handbook of Quantitative Methods. Statistical Analyses* 2 (2013). Publisher: Oxford University Press, pp. 142–153.

[67] Panos P Markopoulos et al. "Efficient L1-norm principal-component analysis via bit flipping". In: *IEEE Transactions on Signal Processing* 65.16 (2017). Publisher: IEEE, pp. 4252–4264.

[68] P Loslever, El M Laassel, and JC Angue. "Combined statistical study of joint angles and ground reaction forces using component and multiple correspondence analysis". In: *IEEE Transactions on biomedical engineering* 41.12 (1994). Publisher: IEEE, pp. 1160–1167.

[69] Nasir Saukani and Noor Azina Ismail. "Identifying the components of social capital by categorical principal component analysis (CATPCA)". In: *Social Indicators Research* 141.2 (2019). Publisher: Springer, pp. 631–655.

[70] Li Yang. "Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction". In: *IEEE transactions on pattern analysis and machine intelligence* 30.3 (2008). Publisher: IEEE, pp. 438–450.

[71] John Sammon. "J.(1969). A nonlinear mapping for data structure analysis". In: *IEEE Transactions on Computers. C-18 (5)* (), pp. 401–409.

[72] Anton K Formann. "Constrained latent class models: Theory and applications". In: *British Journal of Mathematical and Statistical Psychology* 38.1 (1985). Publisher: Wiley Online Library, pp. 87–111.

[73] Simon Lacoste-Julien, Fei Sha, and Michael Jordan. "DiscLDA: Discriminative learning for dimensionality reduction and classification". In: *Advances in neural information processing systems* 21 (2008).

[74] Zhihua Zhang and Michael I Jordan. "Latent variable models for dimensionality reduction". In: *Artificial intelligence and statistics*. PMLR, 2009, pp. 655–662.

[75] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[76] Vilfredo Pareto. *Cours d'économie politique*. Vol. 1. Librairie Droz, 1964.

[77] D.W. Opitz and R.F. Maclin. "An empirical evaluation of bagging and boosting for artificial neural networks". In: *Proceedings of International Conference on Neural Networks (ICNN'97)*. Proceedings of International Conference on Neural Networks (ICNN'97). Vol. 3. June 1997, 1401–1405 vol.3. DOI: 10.1109/ICNN.1997.613999.

[78] Vitor Brock and Habib Ullah Khan. "Big data analytics: does organizational factor matters impact technology acceptance?" In: *Journal of Big Data* 4.1 (2017). Publisher: Springer, pp. 1–28.

[79] Yueran Yang. "Evaluating classification performance: Receiver operating characteristic and expected utility." In: *Psychological Methods* (2022). Publisher: American Psychological Association.

[80] Nigam H Shah, Arnold Milstein, and Steven C Bagley. "Making machine learning models clinically useful". In: *Jama* 322.14 (2019), pp. 1351–1352.

[81] Luca Casini et al. "The Barrier of Meaning in Archaeological Data Science". In: *Proceedings SCIFI-IT'2020 - 4th Annual Science Fiction Prototyping Conference*. Ghent: Eurosis, 2020, pp. 61–65. ISBN: 978-94-92859-10-5. URL: http://arxiv.org/abs/2102.06022v1.

[82] M. Roccetti et al. "Potential and Limitations of Designing a Deep Learning Model for Discovering New Archaeological Sites: A Case with the Mesopotamian Floodplain". In: *Proceedings 6th EAI International Conference on Smart Objects and Technologies for Social Good, GOODTECHS 2020*. New York: ACM International Conference Proceeding Series, 2020, pp. 216–221. ISBN: 978-1-4503-7559-7. DOI: 10.1145/3411170.3411254.

[83] L. Casini et al. "When Machines Find Sites for the Archaeologists: A Preliminary Study with Semantic Segmentation applied on Satellite Imagery of the Mesopotamian Floodplain". In: ACM International Conference Proceeding Series. 2022, pp. 378–383. DOI: 10.1145/3524458.3547121.

[84] Masoud Mahdianpari et al. "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery". In: *Remote Sensing* 10.7 (2018). Publisher: MDPI, p. 1119.

[85] Jean-Daniel Sylvain, Guillaume Drolet, and Nicolas Brown. "Mapping dead forest cover using a deep convolutional neural network and digital aerial photography". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 156 (2019). Publisher: Elsevier, pp. 14–26.

[86] John C Bliss, Thomas M Bonnicksen, and Thomas H Mace. "Computer-aided classification of forest cover types from small scale aerial photography". In: *Environmental Management* 4.6 (1980). Publisher: Springer, pp. 499–510.

[87] Zsolt Domozi and Andras Molnar. "Surveying private pools in suburban areas with neural network based on drone photos". In: *IEEE EUROCON 2019-18th International Conference on Smart Technologies*. IEEE, 2019, pp. 1–6.

[88] Carrie Hritz. "Tracing settlement patterns and channel systems in southern Mesopotamia using remote sensing". In: *Journal of Field Archaeology* 35.2 (2010). Publisher: Taylor & Francis, pp. 184–203.

[89] Sam Redfern. "A PC-based system for computer assisted archaeological interpretation of aerial photographs". In: *BAR INTERNATIONAL SERIES* 750 (1999). Publisher: TEMPUS REPARATSM, pp. 162–164.

[90] Daniele Cerra et al. "An objective assessment of hyperspectral indicators for the detection of buried archaeological relics". In: *Remote Sensing* 10.4 (2018). Publisher: MDPI, p. 500.

[91] Stephen H. Savage, Thomas E. Levy, and Ian W. Jones. "Prospects and problems in the use of hyperspectral imagery for archaeological remote sensing: a case study from the Faynan copper mining district, Jordan". In: *Journal of Archaeological Science* 39.2 (2012), pp. 407–420. ISSN: 0305-4403. DOI: https://doi.org/10.1016/j.jas.2011.09.028. URL: https://www.sciencedirect.com/science/article/pii/S0305440311003554.

[92] Jean-Pierre Toumazet et al. "Automatic detection of complex archaeological grazing structures using airborne laser scanning data". In: *Journal of Archaeological Science: Reports* 12 (2017). Publisher: Elsevier, pp. 569–579.

[93] Enrique Cerrillo-Cuenca. "An approach to the automatic surveying of prehistoric barrows through LiDAR". In: *Quaternary International* 435 (2017). Publisher: Elsevier, pp. 135–145.

[94] Miao Li et al. "A Review of Remote Sensing Image Classification Techniques: the Role of Spatio-contextual Information". In: *European Journal of Remote Sensing* 47.1 (Jan. 2014), pp. 389–411. ISSN: 2279-7254. DOI: 10.5721/EuJRS20144723. URL: https://www.tandfonline.com/doi/full/10.5721/EuJRS20144723 (visited on 10/03/2022).

[95] Nicolò Marchetti et al. "The rise of urbanized landscapes in Mesopotamia: The QADIS integrated survey results and the interpretation of multi-layered historical landscapes". In: *Zeitschrift für Assyriologie und vorderasiatische Archäologie* 109.2 (2019). Publisher: De Gruyter, pp. 214–237.

[96] Bjoern H Menze, Simone Mühl, and Andrew G Sherratt. "Virtual survey on north Mesopotamian tell sites by means of satellite remote sensing". In: *Broadening horizons: multidisciplinary approaches to landscape study* (2007), pp. 5–29.

[97] R.M.C. Adams and R.M.C. Adams. *Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates*. University of Chicago Press, 1981. ISBN: 978-0-226-00544-7. URL: https://books.google.it/books?id=JQaNQgAACAAJ.

[98] Xiaobing Han et al. "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification". In: *Remote Sensing* 9.8 (2017). Publisher: Multidisciplinary Digital Publishing Institute, p. 848.

[99] Dimitrios Marmanis et al. "Deep learning earth observation classification using ImageNet pretrained networks". In: *IEEE Geoscience and Remote Sensing Letters* 13.1 (2015). Publisher: IEEE, pp. 105–109.

[100] Francois Chollet. *Deep learning with Python*. Manning Publications, 2018. ISBN: 978-1-61729-443-3. URL: https://cds.cern.ch/record/2301910 (visited on 10/30/2020).

[101] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[102] Giovanni Delnevo, Marco Roccetti, and Silvia Mirri. "Intelligent and good machines? The role of domain and context codification". In: *Mobile Networks and Applications* 25.3 (2020). Publisher: Springer, pp. 977–985.

[103] Jorge Lazo. "Detection of archaeological sites from aerial imagery using deep learning". In: *Master Thesis* (2019). Publisher: Lund University Department of Astronomy and Theoretical Physics. URL: http://lup.lub.lu.se/student-papers/record/8974790.

[104] Hector A. Orengo et al. "Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data". In: *Proceedings of the National Academy of Sciences* 117.31 (Aug. 4, 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 18240–18250. DOI: 10.1073/pnas.2005583117. URL: https://www.pnas.org/doi/10.1073/pnas.2005583117 (visited on 10/18/2022).

[105] Jesse Casana. "Global-Scale Archaeological Prospection using CORONA Satellite Imagery: Automated, Crowd-Sourced, and Expert-led Approaches". In: *Journal of Field Archaeology* 45 (sup1 Feb. 20, 2020). Publisher: Routledge _eprint: https://doi.org/10.1080/00934690.2020.1713285, S89–S100. ISSN: 0093-4690. DOI: 10.1080/00934690.2020.1713285. URL: https://doi.org/10.1080/00934690.2020.1713285 (visited on 10/18/2022).

[106] Alexander Buslaev et al. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (Feb. 2020). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 125. ISSN: 2078-2489. DOI: 10.3390/info11020125. URL: https://www.mdpi.com/2078-2489/11/2/125 (visited on 09/30/2022).

[107] Simone Mantellini and Amriddin E. Berdimuradov. "Evaluating the human impact on the archaeological landscape of Samarkand (Uzbekistan): A diachronic assessment of the Taylak district by remote sensing, field survey, and local knowledge". In: *Archaeological Research in Asia* 20 (Dec. 1, 2019), p. 100143. ISSN: 2352-2267. DOI: 10.1016/j.ara.2019.100143. URL: https://www.sciencedirect.com/science/article/pii/S2352226718300254 (visited on 10/19/2022).

[108] Pavel Yakubovskiy. *Segmentation Models Pytorch*. Publication Title: GitHub repository. 2020. URL: https://github.com/qubvel/segmentation_models.pytorch.

[109] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[110] Tongle Fan et al. "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation". In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 179656–179665. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3025372.

[111] Rui Li et al. "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). Conference Name: IEEE Transactions on

Geoscience and Remote Sensing, pp. 1–13. ISSN: 1558-0644. DOI: 10.1109/TGRS.2021.3093977.

[112] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[113] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[114] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[115] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[116] Ricardo Baeza-Yates and Marina Estévez-Almenzar. "The Relevance of Non-Human Errors in Machine Learning". In: *EBeM'22: Workshop on AI Evaluation Beyond Metrics*. 2022.

[117] Li-Chia Yang and Alexander Lerch. "On the evaluation of generative models in music". In: *Neural Computing and Applications* 32.9 (May 1, 2020), pp. 4773–4784. ISSN: 1433-3058. DOI: 10.1007/s00521-018-3849-7. URL: https://doi.org/10.1007/s00521-018-3849-7 (visited on 10/03/2022).

[118] J. D. Fernandez and F. Vico. "AI Methods in Algorithmic Composition: A Comprehensive Survey". In: *Journal of Artificial Intelligence Research* 48 (Nov. 17, 2013), pp. 513–582. ISSN: 1076-9757. DOI: 10.1613/jair.3908. URL: https://www.jair.org/index.php/jair/article/view/10845 (visited on 10/03/2022).

[119] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep Learning Techniques for Music Generation*. Computational Synthesis and Creative Systems. Cham: Springer International Publishing, 2020. ISBN: 978-3-319-70162-2. DOI: 10.1007/978-3-319-70163-9. URL: http://link.springer.com/10.1007/978-3-319-70163-9 (visited on 10/03/2022).

[120] Luca Casini, Gustavo Marfia, and Marco Roccetti. "Some Reflections on the Potential and Limitations of Deep Learning for Automated Music Generation". In: *Proceedings 29th IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2018, Bologna*. Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc., 2018,

pp. 27–31. ISBN: 978-1-5386-6009-6. DOI: 10.1109/PIMRC.2018.8581038. URL: 10.1109/PIMRC.2018.8581038.

[121] Luca Casini and Bob L. T. Sturm. "Tradformer: A Transformer Model of Traditional Music Transcriptions". In: Thirty-First International Joint Conference on Artificial Intelligence. Vol. 6. ISSN: 1045-0823. July 16, 2022, pp. 4915–4920. DOI: 10.24963/ijcai.2022/681. URL: https://www.ijcai.org/proceedings/2022/681 (visited on 10/05/2022).

[122] B. L. Sturm et al. "Music Transcription Modelling and Composition Using Deep Learning". In: *Proc. Conf. Computer Simulation of Musical Creativity*. Huddersfield, UK, 2016.

[123] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[124] Angelos Katharopoulos et al. "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, Nov. 21, 2020, pp. 5156–5165. URL: https://proceedings.mlr.press/v119/katharopoulos20a.html (visited on 10/05/2022).

[125] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[126] B. L. Sturm. "How Stuff Works: LSTM Model of Folk Music Transcriptions". In: *Proc. Joint Workshop on Machine Learning for Music, ICML*. 2018.

[127] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 11–20.

[128] Ari Holtzman et al. "The Curious Case of Neural Text Degeneration". In: *International Conference on Learning Representations*. 2019.

[129] Uri Shaham and Omer Levy. "What Do You Get When You Cross Beam Search with Nucleus Sampling?" In: *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 38–45. DOI: 10.18653/v1/2022.insights-1.5. URL: https://aclanthology.org/2022.insights-1.5.

[130] Eric Hallström et al. "From Jigs and Reels to Schottisar och Polskor: Generating Scandinavian-like Folk Music with Deep Recurrent Networks". In: *The 16th Sound & Music Computing Conference, Malaga, Spain, 28-31 May 2019*. 2019.

[131] Bob L Sturm and Oded Ben-Tal. "Taking the models back to music practice: Evaluating generative transcription models built using deep learning". In: *Journal of Creative Music Systems* 2 (2017), pp. 32–60.

[132] L. Casini and M. Roccetti. "The impact of AI on the musical world: will musicians be obsolete?" In: *STUDI DI ESTETICA* 12 (2018), pp. 119–134. DOI: 10.7413/18258646064. URL: http://mimesisedizioni.it/journals/index.php/studi-di-estetica/article/view/630.

[133] Minhyang (Mia) Suh et al. "AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–11. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445219. URL: https://doi.org/10.1145/3411764.3445219 (visited on 10/03/2022).

[134] M. Pearce, D. Meredith, and G. Wiggins. "Motivations and methodologies for automation of the compositional process". In: *Musicae Scientiae* 6.2 (2002), pp. 119–147.

[135] L. Casini and M. Roccetti. "A cross-regional analysis of the COVID-19 spread during the 2020 Italian vacation period: Results from three computational models are compared". In: *SENSORS* 20 (2020), pp. 1–22. DOI: 10.3390/s20247319. URL: https://www.mdpi.com/1424-8220/20/24/7319.

[136] L. Casini and M. Roccetti. "Fashion, Digital Technologies, and AI. Is the 2020 Pandemic Really Driving a Paradigm Shift?" In: *ZONEMODA JOURNAL* 10 (2020), pp. 1–10. DOI: 10.6092/issn.2611-0563/11802. URL: https://zmj.unibo.it/article/view/11802.

[137] Luca Casini and Marco Roccetti. "Reopening Italy's schools in September 2020: a Bayesian estimation of the change in the growth rate of new SARS-CoV-2 cases". In: *BMJ Open* 11.7 (July 1, 2021), e051458. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2021-051458. URL: https://bmjopen.bmj.com/content/11/7/e051458 (visited on 09/21/2022).

[138] L. Casini and M. Roccetti. "A bayesian analysis of the inversion of the sars-cov-2 case rate in the countries of the 2020 european football championship". In: *FUTURE INTERNET* 13 (2021), pp. 1–16. DOI: 10.3390/fi13080212.

[139] M. Roccetti, K. A. Velasco, and L. Casini. "The Influence of Atmospheric Particulate on the Second Wave of CoViD-19 Pandemic in Emilia-Romagna (Italy): Some Empirical Findings". In: *Lecture Notes in Networks and Systems*. Vol. 319. Springer Science and Business Media Deutschland GmbH,

2022, pp. 983–988. ISBN: 978-3-030-85540-6. DOI: 10.1007/978-3-030-85540-6_125.

[140] M. Roccetti and L. Casini. "The role of inter-regional tourism in the spread of COVID-19 in Italy during the 2020 Summer: A confirmatory study". In: *GoodIT 2021 - Proceedings of the 2021 Conference on Information Technology for Social Good*. Association for Computing Machinery, Inc, 2021, pp. 1–6. ISBN: 978-1-4503-8478-0. DOI: 10.1145/3462203.3475888.

[141] R. Cappi et al. "Questioning the seasonality of SARS-COV-2: a Fourier spectral analysis". In: *BMJ OPEN* 12 (2022), pp. 1–12. DOI: 10.1136/bmjopen-2022-061602. URL: https://bmjopen.bmj.com/content/12/4/e061602.

[142] Umang Bhatt et al. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 401–413. ISBN: 9781450384735. DOI: 10.1145/3461702.3462571. URL: https://doi.org/10.1145/3461702.3462571.

[143] Meg Miller. "2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository: Johns Hopkins University Center for Systems Science and Engineering". In: *Bulletin - Association of Canadian Map Libraries and Archives (ACMLA)* 164 (Mar. 30, 2020), pp. 47–51. ISSN: 2561-2263. DOI: 10.15353/acmla.n164.1730. URL: https://openjournals.uwaterloo.ca/index.php/acmla/article/view/1730 (visited on 09/21/2022).

[144] Wladimir Koppen. "Das geographische system der klimat". In: *Handbuch der klimatologie* (1936). Publisher: Gebruder Borntraeger, p. 46.

[145] Yahua Zhang, Anming Zhang, and Jiaoe Wang. "Exploring the roles of high-speed train, air and coach services in the spread of COVID-19 in China". In: *Transport Policy* 94 (Aug. 1, 2020), pp. 34–42. ISSN: 0967-070X. DOI: 10.1016/j.tranpol.2020.05.012. URL: http://www.sciencedirect.com/science/article/pii/S0967070X20304273 (visited on 10/26/2020).

[146] Tomás Pueyo, Nathaniel Lash, and Yaryna Serkez. "This Is Why We Couldn't Control the Pandemic". In: *The New York Times* (Sept. 14, 2020). ISSN: 0362-4331. URL: https://www.nytimes.com/interactive/2020/09/14/opinion/politics/coronavirus-close-borders-travel-quarantine.html (visited on 10/26/2020).

[147] Presidenza del Consiglio dei Ministri. "Decreto del Presidente del Consiglio dei Ministri, 8 Marzo". In: *Gazzetta Ufficiale della Repubblica Italiana* Gazzetta Ufficiale della Repubblica Italiana (Aug. 3, 2020). URL: https://www.gazzettaufficiale.it/eli/id/2020/03/08/20A01522/sg (visited on 10/30/2020).

[148] Presidenza del Consiglio dei Ministri. "Decreto del Presidente del Consiglio dei Ministri, 16 Maggio". In: *Gazzetta Ufficiale della Repubblica Italiana* (May 16, 2020). URL: https://www.gazzettaufficiale.it/eli/id/2020/05/16/20G00051/sg (visited on 10/30/2020).

[149] Civil Protection Department. *COVID-19 Italian Data Repository*. 2020. URL: https://github.com/pcm-dpc/COVID-19 (visited on 10/30/2020).

[150] Virginia Pietromarchi. *Italy's busy summer lights fuse on coronavirus resurgence fears*. Aug. 28, 2020. URL: https://www.aljazeera.com/news/2020/8/28/italys-busy-summer-lights-fuse-on-coronavirus-resurgence-fears (visited on 10/30/2020).

[151] Lindsay Matthews. "Italy Reopens to European Travelers—but Not to Americans Yet". In: *AFAR Media* (Mar. 6, 2020). URL: https://www.afar.com/magazine/is-italy-reopening-and-when-will-i-be-able-to-visit (visited on 10/30/2020).

[152] Angela Giuffrida. "How Sardinia went from safe haven to Covid-19 hotspot". In: *The Guardian* (Sept. 6, 2020). ISSN: 0261-3077. URL: https://www.theguardian.com/world/2020/sep/06/how-sardinia-went-from-safe-haven-to-covid-19-hotspot (visited on 10/30/2020).

[153] Hiroyuki Furuya. "Risk of transmission of airborne infection during train commute based on mathematical model". In: *Environmental Health and Preventive Medicine* 12.2 (Mar. 2007), pp. 78–83. ISSN: 1342-078X. DOI: 10.1007/BF02898153. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723643/ (visited on 10/26/2020).

[154] Maogui Hu et al. "Risk of Coronavirus Disease 2019 Transmission in Train Passengers: an Epidemiological and Modeling Study". In: *Clinical Infectious Diseases* (2020). DOI: 10.1093/cid/ciaa1057. URL: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1057/5877944 (visited on 10/26/2020).

[155] Tamás Krisztin, Philipp Piribauer, and Michael Wögerer. "The spatial econometrics of the coronavirus pandemic". In: *Letters in Spatial and Resource Sciences* (Aug. 1, 2020). ISSN: 1864-404X. DOI: 10.1007/s12076-020-00254-1. URL: https://doi.org/10.1007/s12076-020-00254-1 (visited on 10/26/2020).

[156] Susmita Dasgupta and David Wheeler. *Modeling and Predicting the Spread of Covid-19: Comparative Results for the United States, the Philippines, and South Africa*. World Bank, Washington, DC, 2020, p. 24.

[157] Mohammad Reza Farzanegan et al. "International Tourism and Outbreak of Coronavirus (COVID-19): A Cross-Country Analysis". In: *Journal of Travel Research* (July 3, 2020), p. 0047287520931593. ISSN: 0047-2875. DOI: 10.1177/0047287520931593. URL: https://doi.org/10.1177/0047287520931593 (visited on 10/26/2020).

[158] Martin Thomas Falk and Eva Hagsten. "The unwanted free rider: Covid-19". In: *Current Issues in Tourism* 0.0 (May 26, 2020), pp. 1–6. ISSN: 1368-3500. DOI: 10.1080/13683500.2020.1769575. URL: https://doi.org/10.1080/13683500.2020.1769575 (visited on 10/26/2020).

[159] Stefan Gössling, Daniel Scott, and C. Michael Hall. "Pandemics, tourism and global change: a rapid assessment of COVID-19". In: *Journal of Sustainable Tourism* 29.1 (Jan. 2, 2021), pp. 1–20. ISSN: 0966-9582. DOI: 10.1080/09669582.2020.1758708. URL: https://doi.org/10.1080/09669582.2020.1758708 (visited on 10/26/2020).

[160] Marco D'Orazio, Gabriele Bernardini, and Enrico Quagliarini. "Sustainable and resilient strategies for touristic cities against COVID-19: An agent-based approach". In: *Safety Science* 142 (Oct. 1, 2021), p. 105399. ISSN: 0925-7535. DOI: 10.1016/j.ssci.2021.105399. URL: https://www.sciencedirect.com/science/article/pii/S0925753521002435 (visited on 09/21/2022).

[161] Weston C. Roda et al. "Why is it difficult to accurately predict the COVID-19 epidemic?" In: *Infectious Disease Modelling* 5 (2020), pp. 271–281. ISSN: 2468-2152. DOI: 10.1016/j.idm.2020.03.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104073/ (visited on 10/26/2020).

[162] G. Rigaill, E. Lebarbier, and S. Robin. "Exact posterior distributions and model selection criteria for multiple change-point detection problems". In: *Statistics and Computing* 22.4 (July 1, 2012), pp. 917–929. ISSN: 1573-1375. DOI: 10.1007/s11222-011-9258-8. URL: https://link.springer.com/article/10.1007/s11222-011-9258-8 (visited on 10/30/2020).

[163] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods". In: *Signal Processing* 167 (Feb. 1, 2020), p. 107299. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2019.107299. URL: http://www.sciencedirect.com/science/article/pii/S0165168419303494 (visited on 10/30/2020).

[164] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Proceedings of the 9th Python in Science Conference* (2010), pp. 92–96. DOI: 10.25080/Majora-92bf1922-011. URL: http://conference.scipy.org/proceedings/scipy2010/seabold.html (visited on 10/30/2020).

[165] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. 4th ed. Statistics and Computing. New York: Springer-Verlag, 2002. ISBN: 978-0-387-95457-8. DOI: 10.1007/978-0-387-21706-2. URL: https://www.springer.com/gp/book/9780387954578 (visited on 10/30/2020).

[166] Max S. Y. Lau et al. "Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA". In: *Proceedings of the National Academy of Sciences* 117.36 (Sept. 8, 2020), pp. 22430–22435. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2011802117. URL: https://www.pnas.org/content/117/36/22430 (visited on 11/09/2020).

[167] J. O. Lloyd-Smith et al. "Superspreading and the effect of individual variation on disease emergence". In: *Nature* 438.7066 (Nov. 2005), pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature04153. URL: https://www.nature.com/articles/nature04153 (visited on 11/09/2020).

[168] Chen Stein-Zamir et al. "A large COVID-19 outbreak in a high school 10 days after schools' reopening, Israel, May 2020". In: *Eurosurveillance* 25.29 (2020). Publisher: European Centre for Disease Prevention and Control, p. 2001352.

[169] Meira Levinson, Muge Cevik, and Marc Lipsitch. *Reopening primary schools during the pandemic*. Issue: 10 Pages: 981–985 Publication Title: New England Journal of Medicine Volume: 383. 2020.

[170] Sharif A Ismail et al. "SARS-CoV-2 infection and transmission in educational settings: a prospective, cross-sectional analysis of infection clusters and outbreaks in England". In: *The Lancet Infectious Diseases* 21.3 (2021). Publisher: Elsevier, pp. 344–353.

[171] Laura Heavey et al. "No evidence of secondary transmission of COVID-19 from children attending school in Ireland, 2020". In: *Eurosurveillance* 25.21 (2020). Publisher: European Centre for Disease Prevention and Control, p. 2000903.

[172] Kristine Macartney et al. "Transmission of SARS-CoV-2 in Australian educational settings: a prospective cohort study". In: *The Lancet Child & Adolescent Health* 4.11 (2020). Publisher: Elsevier, pp. 807–816.

[173] Henrik Salje et al. "Estimating the burden of SARS-CoV-2 in France". In: *Science* 369.6500 (2020). Publisher: American Association for the Advancement of Science, pp. 208–211.

[174] Elizabeth M Dufort et al. "Multisystem inflammatory syndrome in children in New York State". In: *New England Journal of Medicine* 383.4 (2020). Publisher: Mass Medical Soc, pp. 347–358.

[175] Leora R Feldstein et al. "Multisystem inflammatory syndrome in US children and adolescents". In: *New England Journal of Medicine* 383.4 (2020). Publisher: Mass Medical Soc, pp. 334–346.

[176] Ramanan Laxminarayan et al. "Epidemiology and transmission dynamics of COVID-19 in two Indian states". In: *Science* 370.6517 (2020). Publisher: American Association for the Advancement of Science, pp. 691–697.

[177] Alasdair PS Munro and Saul N Faust. "Children are not COVID-19 super spreaders: time to go back to school". In: *Archives of disease in childhood* 105.7 (2020). Publisher: BMJ Publishing Group Ltd, pp. 618–619.

[178] Russell M Viner et al. "Susceptibility to SARS-CoV-2 infection among children and adolescents compared with adults: a systematic review and meta-analysis". In: *JAMA pediatrics* 175.2 (2021). Publisher: American Medical Association, pp. 143–156.

[179] Jonas F Ludvigsson. "Children are unlikely to be the main drivers of the COVID-19 pandemic–a systematic review". In: *Acta Paediatrica* 109.8 (2020). Publisher: Wiley Online Library, pp. 1525–1530.

[180] Shao-Yi Cheng et al. "How to safely reopen colleges and universities during COVID-19: experiences from Taiwan". In: *Annals of internal medicine* 173.8 (2020). Publisher: American College of Physicians, pp. 638–641.

[181] Steven Riley et al. "High prevalence of SARS-CoV-2 swab positivity and increasing R number in England during October 2020: REACT-1 round 6 interim report". In: *MedRxiv* (2020). Publisher: Cold Spring Harbor Laboratory Press.

[182] UK Office for National Statistics. *Coronavirus (COVID-19) infection survey*. Version Number: 2022-11-13. UK Office for National Statistics, Nov. 13, 2022.

[183] Stefan Flasche and W John Edmunds. "The role of schools and school-aged children in SARS-CoV-2 transmission". In: *The Lancet infectious diseases* 21.3 (2021). Publisher: Elsevier, pp. 298–299.

[184] Gavin Yamey and Rochelle P Walensky. "Covid-19: re-opening universities is high risk". In: *BMJ* (Sept. 1, 2020), p. m3365. ISSN: 1756-1833. DOI: 10.1136/bmj.m3365. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.m3365 (visited on 10/04/2022).

[185] Giovanni Sebastiani and Giorgio Palù. "COVID-19 and School Activities in Italy". In: *Viruses* 12.11 (Nov. 2020), p. 1339. ISSN: 1999-4915. DOI: 10.3390/v12111339. URL: https://www.mdpi.com/1999-4915/12/11/1339 (visited on 09/20/2022).

[186] Elisabetta Larosa et al. "Secondary transmission of COVID-19 in preschool and school settings in northern Italy after their reopening in September 2020: a population-based study". In: *Eurosurveillance* 25.49 (2020). Publisher: European Centre for Disease Prevention and Control, p. 2001911.

[187] Crispian Balmer and Antonio Denti. "Happiness, controversy and fear as Italy's schools finally reopen". In: *Reuters* (Sept. 14, 2020). URL: https://www.reuters.com/article/us-health-coronavirus-italy-schools-idUSKBN26519H (visited on 09/19/2022).

[188] Perri Klass. "Back-to-School Season in Italy". In: *The New York Times* (Sept. 28, 2020). ISSN: 0362-4331. URL: https://www.nytimes.com/2020/09/28/well/family/back-to-school-season-in-italy.html (visited on 09/19/2022).

[189] Danilo Buonsenso et al. "SARS-CoV-2 Infections in Italian Schools: Preliminary Findings After 1 Month of School Opening During the Second Wave of the Pandemic". In: *Frontiers in Pediatrics* 8 (2021). ISSN: 2296-2360. URL: https://www.frontiersin.org/articles/10.3389/fped.2020.615894 (visited on 09/19/2022).

[190] D. A. Stephens. "Bayesian Retrospective Multiple-Changepoint Identification". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1 (1994). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/2986119, pp. 159–178. ISSN: 1467-9876. DOI: 10.2307/2986119. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/2986119 (visited on 09/21/2022).

[191] Jonas K Lindeløv. *mcp: An R Package for Regression With Multiple Change Points*. Jan. 2020. DOI: 10.31219/osf.io/fzqxv. URL: osf.io/fzqxv.

[192] Google. *Coronavirus (COVID-19) - Google mobility Reports*. Google mobility Reports. June 10, 2020. URL: https://www.gstatic.com/covid19/mobility/2020-10-06_IT_Mobility_Report_it.pdf (visited on 10/06/2020).

[193] Mara Sanfelici. "The Italian response to the COVID-19 crisis: lessons learned and future direction in social development". In: *The International Journal of Community and Social Development* 2.2 (2020). Publisher: SAGE Publications Sage India: New Delhi, India, pp. 191–210.

[194] Emma Thomasson. "German minister chides 'irresponsible' UEFA over Euro 2020 crowds". In: *Reuters* (July 1, 2021). URL: https://www.reuters.com/world/europe/german-minister-slams-uefas-decision-fuller-stadiums-2021-07-01/ (visited on 09/20/2022).

[195] World Health Organization. *COVID-19: The stakes are still high*. World Health Organization. July 1, 2021. URL: https://www.who.int/europe/news/item/01-07-2021-covid-19-the-stakes-are-still-high (visited on 09/20/2022).

[196] Jon Henley and Jennifer Rankin. "Covid: Euro 2020 crowds 'a recipe for disaster', warns EU committee". In: *The Guardian* (July 1, 2021). ISSN: 0261-3077. URL: https://www.theguardian.com/world/2021/jul/01/covid-euro-2020-crowds-a-recipe-for-disaster-warns-german-minister (visited on 09/20/2022).

[197] UEFA.com. *Key information for EURO 2020 spectators*. Apr. 23, 2021. URL: https://www.uefa.com/uefaeuro/history/news/025b-0ef33753d7d0-100629325be2-1000--key-information-for-euro-spectators/ (visited on 09/20/2022).

[198] Agenzia Nazionale Stampa Associata. *Cluster of 91 COVID-19 cases linked to Euro 2020 game - English*. ANSA.it. July 16, 2021. URL: https://www.ansa.it/english/news/general_news/2021/07/16/cluster-of-91-covid-19-cases-linked-to-euro-2020-game_84349124-e130-453b-ade2-8b7136bd8993.html (visited on 09/20/2022).

[199] Tom Kington. "Italy's Euro 2020 victory tour sent Rome cases rocketing". In: *The Times* (July 22, 2021). ISSN: 0140-0460. URL: https://www.thetimes.co.uk/article/italys-euro-2020-victory-tour-sent-rome-cases-rocketing-r6m667r0b (visited on 09/20/2022).

[200] Elian Peltier. "Crowds for European Championship soccer games are driving infections, the W.H.O. says." In: *The New York Times* (July 1, 2021). ISSN: 0362-4331. URL: https://www.nytimes.com/2021/07/01/world/europe/euro-2020-covid-outbreak.html (visited on 09/20/2022).

[201] Nikolaj Skydsgaard and Jacob Gronholt-pedersen. "Euro 2020 crowds driving rise in COVID-19 infections, says WHO". In: *Reuters* (July 1, 2021). URL: https://www.reuters.com/world/europe/who-warns-third-coronavirus-wave-europe-2021-07-01/ (visited on 09/20/2022).

[202] Yorck Olaf Schumacher et al. "Resuming professional football (soccer) during the COVID-19 pandemic in a country with high infection rates: a prospective cohort study". In: *British Journal of Sports Medicine* 55.19 (Oct. 1, 2021). Publisher: BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine Section: Original research, pp. 1092–1098. ISSN: 0306-3674, 1473-0480. DOI: 10.1136/bjsports-2020-103724. URL: https://bjsm.bmj.com/content/55/19/1092 (visited on 09/21/2022).

[203] Marek Kochańczyk, Frederic Grabowski, and Tomasz Lipniacki. "Super-spreading events initiated the exponential growth phase of COVID-19 with R0 higher than initially estimated". In: *Royal Society Open Science* 7.9 (). Publisher: Royal Society, p. 200786. DOI: 10.1098/rsos.200786. URL: https://royalsocietypublishing.org/doi/full/10.1098/rsos.200786 (visited on 09/21/2022).

[204] Mike Weed and Abby Foad. *Rapid Scoping Review of Evidence of Outdoor Transmission of COVID-19*. Sept. 10, 2020. DOI: 10.1101/2020.09.04.20188417. URL: https://www.medrxiv.org/content/10.1101/2020.09.04.20188417v2 (visited on 09/20/2022).

[205] D. Cereda et al. *The early phase of the COVID-19 outbreak in Lombardy, Italy*. Issue: arXiv:2003.09320. Mar. 20, 2020. DOI: 10.48550/arXiv.2003.09320. arXiv: 2003.09320[q-bio]. URL: http://arxiv.org/abs/2003.09320 (visited on 09/20/2022).

[206] Moritz Mercker, Uwe Betzin, and Dennis Wilken. *What influences COVID-19 infection rates: A statistical approach to identify promising factors applied to infection data from Germany*. Apr. 17, 2020. DOI: 10.1101/2020.04.14.20064501. URL: https://www.medrxiv.org/content/10.1101/2020.04.14.20064501v1 (visited on 09/20/2022).

[207] Carlo Signorelli et al. "Major sports events and the transmission of SARS-CoV-2: analysis of seven case-studies in Europe". In: *Acta Bio Medica : Atenei Parmensis* 91.2 (2020), pp. 242–244. ISSN: 0392-4203. DOI: 10.23750/abm.v91i2.9699. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7569655/ (visited on 09/21/2022).

[208] Deloitte. *Football Money League clubs will miss out on revenue of over €2 billion - Deloitte Middle East Press release*. Jan. 27, 2021. URL: https://www2.deloitte.com/xe/en/pages/about-deloitte/articles/deloittes-sports-business-group-estimates-football-money-league-miss-out-revenue-over-2euros-billion-end-202021-due-covid.html (visited on 09/20/2022).

[209] Filippo D'Amico et al. "COVID-19 seasonality in temperate countries". In: *Environmental Research* 206 (Apr. 15, 2022), p. 112614. ISSN: 0013-9351. DOI: 10.1016/j.envres.2021.112614. URL: https://www.sciencedirect.com/science/article/pii/S0013935121019150 (visited on 09/21/2022).

[210] Yiqun Ma et al. "Role of meteorological factors in the transmission of SARS-CoV-2 in the United States". In: *Nature Communications* 12.1 (June 14, 2021), p. 3602. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23866-7. URL: https://www.nature.com/articles/s41467-021-23866-7 (visited on 09/21/2022).

[211] Alejandro Fontal et al. "Climatic signatures in the different COVID-19 pandemic waves across both hemispheres". In: *Nature Computational Science* 1.10 (Oct. 2021), pp. 655–665. ISSN: 2662-8457. DOI: 10.1038/s43588-021-00136-6. URL: https://www.nature.com/articles/s43588-021-00136-6 (visited on 09/21/2022).

[212] Francesco Sera et al. "A cross-sectional analysis of meteorological factors and SARS-CoV-2 transmission in 409 cities across 26 countries". In: *Nature Communications* 12.1 (Oct. 13, 2021), p. 5968. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25914-8. URL: https://www.nature.com/articles/s41467-021-25914-8 (visited on 09/21/2022).

[213] Rachel E. Baker et al. "Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic". In: *Science* 369.6501 (July 17, 2020), pp. 315–319. DOI: 10.1126/science.abc2535. URL: https://www.science.org/doi/10.1126/science.abc2535 (visited on 09/21/2022).

[214] Charles Roberto Telles, Henrique Lopes, and Diogo Franco. "SARS-COV-2: SIR Model Limitations and Predictive Constraints". In: *Symmetry* 13.4 (Apr. 2021), p. 676. ISSN: 2073-8994. DOI: 10.3390/sym13040676. URL: https://www.mdpi.com/2073-8994/13/4/676 (visited on 09/21/2022).

[215] Yinon M. Bar-On et al. "Protection of BNT162b2 Vaccine Booster against Covid-19 in Israel". In: *New England Journal of Medicine* 385.15 (Oct. 7, 2021), pp. 1393–1400. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2114255. URL: https://www.nejm.org/doi/10.1056/NEJMoa2114255 (visited on 09/21/2022).

[216] Takumi Kato. "Opposition in Japan to the Olympics during the COVID-19 pandemic". In: *Humanities and Social Sciences Communications* 8.1 (Dec. 16, 2021), pp. 1–9. ISSN: 2662-9992. DOI: 10.1057/s41599-021-01011-5. URL: https://www.nature.com/articles/s41599-021-01011-5 (visited on 09/21/2022).

[217] Jennifer Jacobs and Sophia Cai. "Biden Declares Success in Beating Pandemic in July 4 Speech". In: *Bloomberg.com* (July 4, 2021). URL: https://www.bloomberg.com/news/articles/2021-07-04/biden-to-appeal-for-vaccinations-after-u-s-missed-july-4-target (visited on 09/21/2022).

[218] Yoshiyasu Takefuji. "Fourier analysis using the number of COVID-19 daily deaths in the US". In: *Epidemiology & Infection* 149 (2021), e64. ISSN: 0950-2688, 1469-4409. DOI: 10.1017/S0950268821000522. URL: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/fourier-analysis-using-the-number-of-covid19-daily-deaths-in-the-us/37FB15CCE2D2D64C85EE9A0C3C95421E (visited on 09/21/2022).

[219] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2. URL: https://www.nature.com/articles/s41592-019-0686-2.

[220] Bernard Cazelles et al. "Time-dependent spectral analysis of epidemiological time-series with wavelets". In: *Journal of The Royal Society Interface* 4.15 (Aug. 22, 2007), pp. 625–636. DOI: 10.1098/rsif.2007.0212. URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2007.0212 (visited on 09/21/2022).

[221] Stephen X Zhang et al. "A Second Wave? What Do People Mean by COVID Waves? – A Working Definition of Epidemic Waves". In: *Risk Management and Healthcare Policy* Volume 14 (Sept. 2021), pp. 3775–3782. ISSN: 1179-1594. DOI: 10.2147/RMHP.S326051. URL: https://www.dovepress.com/a-second-wave-what-do-people-mean-by-covid-waves--a-working-definition-peer-reviewed-fulltext-article-RMHP (visited on 09/21/2022).

[222] Emma B. Hodcroft et al. "Spread of a SARS-CoV-2 variant through Europe in the summer of 2020". In: *Nature* 595.7869 (July 2021), pp. 707–712. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03677-y. URL: https://www.nature.com/articles/s41586-021-03677-y (visited on 09/21/2022).

[223] James Hadfield et al. "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (Dec. 1, 2018), pp. 4121–4123. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty407. URL: https://doi.org/10.1093/bioinformatics/bty407 (visited on 09/21/2022).

[224] Stefan Elbe and Gemma Buckland-Merrett. "Data, disease and diplomacy: GISAID's innovative contribution to global health". In: *Global Challenges* 1.1 (2017), pp. 33–46. ISSN: 2056-6646. DOI: 10.1002/gch2.1018. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.1018 (visited on 09/21/2022).