Masters Program in Geospatial Technologies

INVESTIGATION OF GEOTHERMAL POTENTIAL WITH MACHINE LEARNING IN MAINLAND PORTUGAL

Matheus Lopes do Nascimento

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

INVESTIGATION OF GEOTHERMAL POTENTIAL WITH MACHINE LEARNING IN MAINLAND PORTUGAL

Dissertation supervised by

Roberto Henriques, PhD
NOVA Information Management School
Universidade Nova de Lisboa
Lisbon, Portugal

and co-supervised by

Vicente de Azevedo Tang, MSc
NOVA Information Management School
Universidade Nova de Lisboa
Lisbon, Portugal

Filiberto Pla Bañón, PhD
Institute of New Imaging Technologies (INIT),
Universitat Jaume I (UJI),
Castellon de la Plana, Spain

February of 2022

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, February 20, 2022

Matheus Lopes do Nascimento

[the signed original has been archived by the NOVA IMS services]

Acknowledgments

First, I would like to thank my supervisors, Dr Roberto Henriques and MSc Vicente Tang at Nova Information Management school, as well as Dr Filiberto Pla at Jaume I University for their guidance and support throughout the thesis research. A special thanks to Dr Elsa Ramalho and Portugal's Geology and Energy Laboratory for the interest in my research and for providing me the data from the Geothermal Atlas of Portugal.

I also want to thank Professor Marco Painho for all advice and support during the master studies. I am very grateful to the Master of Science in Geospatial Technologies program for the opportunity and amazing experience I was given.

To all Universities, professors and staff, that helped me during this period and inspired me to learn and develop new skills and knowledge.

To all the friends I have made in this program for the all the moments we have been through, the friendship, motivation, and support.

Lastly, I would like to express all my love and gratitude to my family for always giving me full support and encouraging me to pursue my goals and decisions in life.

INESTIGATION OF GEOTHERMAL POTENTIAL WITH MACHINE LEARNING IN MAINLAND PORTUGAL

ABSTRACT

Earth's internal heat is explored to produce electricity or used directly in industrial processes or residencies. It is considered to be renewable and cleaner than fossil fuels and has great importance to pursue environmental goals. The exploration phase of geothermal resources is complex and expensive. It requires field surveys, geological, geophysical and geochemical analysis, as well as drilling campaigns. Geospatial data and technologies have been used to target promising sites for further investigations, and helped reduce costs while also pointed to important criteria data related to geothermal potential. Machine learning is a data driven set of technologies that has been successfully used to model environmental parameters, and in the field of geothermal energy it has been used to predict thermal properties of the surface and subsurface. Random Forests and Extreme Gradient Boosting are ensemble machine learning algorithms that have been extensively used in environmental and geological sciences, and have been demonstrated to perform well when predicting thermal properties. This study investigated a methodology that coupled GIS and ML to predict two crucial parameters in geothermal exploration throughout Mainland Portugal: Geothermal gradient and surface Heat flow density. Training data consisted in different types of wells drilled in the study area where the two labels were measured. It was provided by Portugal's Geology and Energy Laboratory. Features were all publicly available and consisted in geological, hydrogeological, geophysical, weather and terrain data. Data were aggregated in two grids with two spatial resolutions. The results between the two algorithms have been compared and discussed. The most important features that contributed to the models were identified and their relationships with the outputs discussed. The models and the prediction maps over the study area showed the location of zones with higher geothermal gradient and surface heat flow density and can be used to aid geothermal exploration and provide insights for geothermal modelling.

KEYWORDS

Geographical Information Systems

Spatial analysis

Machine Learning

Random Forests

Extreme Gradient Boosting

Geothermal exploration

Geothermal gradient

Heat flow

ACRONYMS

AHP – Analytical Hierarchical Process

DEM – Digital Elevation Model

EGS – Enhanced Geothermal Systems

GHF – Geothermal Heat Flux

GIS – Geographical Information Systems

IRENA – International Renewable Energy Agency

LNEG – Laboratório Nacional de Energia e Geologia (Portugal's Geology and Energy Laboratory

MCDA – Multi Criteria Decision Analysis

ML – Machine Learning

QAFI – Quaternary Faults Database of Iberia

RF – Random Forests

XGB – Extreme Gradient Boosting

WFS – Web Feature Service

INDEX OF THE TEXT

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

Geothermal energy is generated by exploring heat stored inside the Earth's crust. The internal heat originated from the processes that formed the planet and is currently sustained by radioactive decay of elements and solar radiation (Barbier, 2002). Geothermal reservoirs are relatively constant, being natural sources of energy that can be used to generate electricity or directly, such as in heating of buildings and industrial processes (Lund & Freeston, 2001). The increasing concern regarding carbon dioxide emissions and global warming has prompted the public and the private sectors to create policies and technologies aimed at developing and improving cleaner and renewable sources of energy production. The energy of the subsurface is considered to be sustainable, and a promising ally in the transition from fossil fuels into a carbon-free energy matrix (Mock et al., 1997).

The exploration stage of geothermal resources is searching for areas that can provide energy or heat to be used economically. Finding sites that have a high potential for implementing a geothermal power plant or for the direct use of the reservoirs requires many resources and it is a demanding task. It is traditionally achieved through in situ investigations and field surveys, such as drilling campaigns, as well as geophysical, geological and geochemical analysis (Barbier, 2002).

Higher geothermal gradients provide higher temperatures at lower depths. However, there are many factors to be taken into account to define if an area is indeed prone to drilling and production. The presence of hot groundwater is only available depending on geological and hydrogeological parameters, such as permeability and stratigraphy of the subsurface (Corniello A., 2012; Suwai, 2009). The feasibility of drilling depends on the mechanical properties of the ground, and the depth in which the temperature is high enough to be explored economically. Traditionally, geothermal energy comes from high-temperature hydrothermal systems, but there are also low-temperature systems, commonly used for direct use of hot water. More recently, geothermal energy is also being produced by fracturing rocks and injecting fluids at locations where water is unavailable or at environments of low permeability, enabling the exploration of geothermal resources in more places whenever there are higher temperature gradients. This process is known as Enhanced Geothermal Systems (EGS) (Olasolo et al., 2016;

Tester, 2007). With the development of EGS, knowing the geothermal gradient and other thermal characteristics of the crust became even more crucial, since it is now possible to develop geothermal power plants regardless of the presence of reachable hot groundwater.

In Mainland Portugal, most of the geothermal resources are low enthalpy systems, exploited primarily for direct use. However, there have been efforts to improve the knowledge about the thermal properties and foment possible future geothermal explorations. With new technologies, such as EGS, this became even more promising and important for an environment with such characteristics. Within these efforts, Portugal's Geology and Energy Laboratory (LNEG) have developed the Geothermal Atlas of Mainland Portugal (Ramalho, 2014, Ramalho & Correia, 2015). The Atlas contains a dataset with geological information and maps created by interpolation methods, built using temperature measurements coming from different types of wells, and by associating the geological formations' thermal conductivity to produce heat flow values. Chamorro et al., (2014) have estimated the theoretical EGS potential in the Iberian Peninsula and verified the presence of thermal energy stored at depths between 3 and 10 km, in a magnitude that is almost five times higher than the current electricity capacity installed in the Peninsula.

Developing new methods to reduce costs and improve efficiency in the exploration phase is a necessary task in this field. For this purpose, studies have been published with successful methodologies using Geographical Information Systems (GIS) by performing spatial analysis, geo statistics, and more recently, Machine Learning (ML) algorithms (Arola et al., 2019; Assouline et al., 2019; Tinti et al., 2018). Most methodologies consisted in either creating suitability indexes or producing maps estimating important mechanical and thermal parameters of the surface or subsurface, which allows for better targeting of field surveys by selecting promising regions regarding geothermal potential.

1.1 Aim and objectives

Taking into consideration the proven capacity of extracting information from geospatial data for many applications in the field of geosciences and geothermal energy, this research couples GIS and ML to assess the potential for geothermal energy exploration

by predicting two crucial parameters of the ground: geothermal gradient and the surface heat flow density.

The aim of this work is to evaluate how accurate can these variables be predicted using publicly available data on geological, hydrogeological, terrain, weather, and geophysical information with the aid of ML models (Random Forests and Extreme Gradient Boosting). These algorithms were chosen due to their proven capacity in geosciences (Merembayev et al., 2019; Sun et al., 2020; Zhang & Zhan, 2017) and in geothermal assessment applications (Assouline et al., 2019; González & Rodríguez-Gonzálvez, 2019; Shahdi et al., 2021). In addition to assessing their accuracy, this study compares the results from the two models, and identifies promising features to predict labels and interprets possible relationships among them.

Two grids with different scales were tested, in order to assess the matter of resolution in the study. In addition, the most important features were identified to provide insights into modelling thermal properties. The dataset and methodology were discussed throughout the literature review and methodology sections. Maps were created at the end, exhibiting the spatial distribution of the predicted geothermal gradient and heat flow density throughout Mainland Portugal. The results and accuracies were demonstrated and discussed. The locations of higher potential areas were overviewed, as well as their correlation to the features used in this research.

2. LITERATURE REVIEW

2.1 Geothermal potential

The potential to explore geothermal energy has been assessed worldwide at both regional and global scales (Assouline et al., 2019; Coro & Trumpy, 2020). It is commonly estimated by the calculation of important physical parameters of the ground and the subsurface, or by the usage of different weighted criteria data, to rank areas regarding their potentiality. The techniques that have been utilized to estimate ground thermal parameters or suitability of areas regarding geothermal resources exploration are varied, with methods changing based on the parameters available and their measurements procedures, as well as the technologies and methods performed to extract information from the available data.

GIS-based spatial analysis has been performed either to create or to extract more information from geospatial data, directly or indirectly associated with geothermal activity. For instance, GIS-based multi criteria decision analysis (MCDA) has been explored as a solution by many authors (Fahil et al., 2020; Omwenga, 2020; Tinti et al., 2018). In this type of research, different geospatial data are weighted based on their relevance to the target and later overlayed to estimate and classify the spatial distribution of geothermal energy exploration suitability classes. In the literature, methods using geo statistics are often applied to interpolate important variables and generate input data for multi criteria decision analysis or ML methodologies. The latter has seen a significant increase in studies focused on this type of methodology to explore geothermal energy, in which studies have been conducted to predict temperatures at specific depths, as well as variables of geothermal gradient, heat flux, thermal conductivity, and other physical properties (Han et al., 2019; Rezvanbehbahani et al., 2017; Shahdi et al., 2021).

The following sections of this chapter cover examples of the aforementioned techniques that have been applied in the literature, while it also bring forward the data sources that have been successfully linked to geothermal potential estimation.

2.2 GIS-based techniques

Analytical Hierarchical Process (AHP) was used in a GIS-based MCDA methodology to delimitate feasible areas for geothermal energy exploration at shallow depths through ground source heat pumps (Tinti et al., 2018). The main parameters utilized were shallow geology, hydrogeology, thermal conductivity and diffusivity of the ground, ambient average temperature, and mechanical parameters such as rock formation's shear strength, to account for the drilling potential. Another MCDA study was performed utilizing a combination of geospatial data such as geological, geochemical and geophysical, in addition to remote sensing images, to orientate geothermal wells siting and targeting at the Eburru geothermal field in Kenya (Omwenga, 2020). Trumpy et al., (2015) used the weighted overlay method combining also thermal, geological, geochemical and geophysical data sources to create favorability maps in Sicily, Italy. Throughout this study, data were interpolated using geo statistics to estimate potential where there were not available data. In the Gulf of Suez, Egypt, the suitability for geothermal exploration was estimated using geological and thermal properties

measured at well loggings and also with remote sensing images, in another weighted overlay methodology (Fahil et al., 2020). Some of the remote sensing derived products were elevation, the density of drainage, and land surface temperature.

These studies have all been successful in identifying zones with higher probability to find explorable geothermal resources and also to establish significant criteria related to the problem. However, their approaches lack the ability to provide estimations of important physical parameters. ML has been tested and applied to generate estimations of target thermal variables, or to pursue better methods to classify favorable areas.

2.3 Machine learning methods

Machine Learning is a data driven technology that has been used to extract information from different sources of data, to predict values for targets when they consist in continuous variables, or to assign classes when they consist in categorial variables (Angra & Ahuja, 2017). It is based on multidisciplinary sets of techniques that cover mathematics, statistics, artificial intelligence, data mining, among others (Kotsiantis et al., 2006). The main objective is to establish relationships between input data (features) and output data (labels). Labels may also be called targets. Training data is fitted into a model in which labels attributes are known. Later, the model is used to predict values by inputting rows containing the same features used for modelling.

In the literature, ML methods have provided efficient and accurate results in several analysis on the exploration and suitability of renewable energies. Shahab & Singh (2019) and Lai et al., (2020) have compared and reviewed distinct ML algorithms being used to classify suitability of areas regarding renewable energy resources. In the field of geothermal energy exploration specifically, ML has been used to predict important thermal properties values, as well as to calculate suitability indexes (Arola et al., 2019; Assouline et al., 2019; Coro & Trumpy, 2020; Gangwani et al., 2020; Mohamed et al., 2015; Shahdi et al., 2021).

A maximum entropy model was used to build a map exhibiting the suitability for the installation of geothermal power plants throughout the globe by using different relevant criteria with global coverage resolutions (Coro & Trumpy, 2020). Shahdi et al., (2021) compared the accuracy and applicability of four different algorithms when predicting subsurface temperature and the geothermal gradient in the Northeast of the United

States. For this task, 20750 wells with bottom hole temperatures and geological information were used. It was concluded that Extreme Gradient Boosting (XGB) and Random Forests (RF), which are both ensemble type of algorithms that work with multiple decision trees, outperformed other algorithms like Neural Networks, and they were also able to generate reliable predictions. Assouline et al., (2019) developed a methodology that used ML, GIS-based data analysis, and physics-based modelling to map the theoretical geothermal potential at shallow depths throughout Switzerland using RF. The research consisted in creating prediction maps of three crucial parameters used to estimate promising geothermal areas, which are: geothermal gradient, ground thermal conductivity and diffusivity.

Arola et al., (2019) used Artificial Neural Networks to build maps of Finland's theoretical potential for geothermal energy at shallow depths, aiming to promote geothermal energy exploration. They assessed the influence of certain types of lithological units and their geographical location with higher geothermal potential, such as granites and quartzites. Another study was performed using Artificial Neural Networks with Bottom Hole Temperatures from oil wells and correlating it to gravity anomalies (Bouguer anomaly) to predict the Geothermal Gradient. In Greenland the Geothermal Heat Flux (GHF) was estimated using a Gradient Boosted Regression Tree (Rezvanbehbahani et al., 2017). The authors were able to establish relationships between lithological and tectonic characteristics to the GHF.

Literature evidenced that ensemble algorithms were extensively used and proven to work efficiently when predicting thermal properties of the surface and subsurface, as well as many other environmental parameters in geosciences. Therefore, it was selected to be further explored in this research.

2.4 Geothermal related variables

Based on the literature review, important features linked to geothermal activity could be identified. This section discusses the data included into this research.

Lithological and structural setting are both correlated to the variables being predicted in this study. Geological faults are structures that can control the flow and pathway when fluids infiltrate the subsurface. This occurs due to the fact that faults can control the permeability in rocks and soils. Geothermal systems were associated with regions

containing a large number of faults and fractures, especially in the presence of deep faults with high permeability (Badino, 2005; Bignall et al., 2010). Permeability of the ground is an important parameter that provides information about fluids flow and infiltration in the subsurface (Corniello A., 2012). The presence of drainages also influences water infiltration and the correlation with geothermal activity increases if the drainage pattern is structurally controlled .

Different types of rock formations and soils have distinct thermal conductivity, that affects the rate that heat is transferred among the surface and subsurface (Ammann et al., 2014; Lerche, 1991; Robertson, 1988). Most aforementioned geothermal studies included at least one level of lithological classification in their analysis.

Bouguer anomaly is a gravity anomaly that allows the identification of the Earth's crust thickness. When the crust is thinner the surface is closer to Earth's mantle and therefore the geothermal gradient is higher. González & Rodríguez-Gonzálvez, (2019); Rezvanbehbahani et al., (2017) have included this type of data into their analysis.

Surface air temperature was included as feature in many cited studies (Coro & Trumpy, 2020; Fahil et al., 2020; González & Rodríguez-Gonzálvez, 2019). It has been correlated to influence thermal energy storage at the subsurface (Nguyen et al., 2017). Local outliers of high temperatures besides lower values can be an indicative of geothermal activity.

Seismic activity is an important factor that can also increase the permeability, due to the presence of active faulting. It was therefore included in this study aiming to establish correlation between the target and different intensity of earthquakes.

Soil texture influences permeability at the surface and its composition can provide information about thermal conductivity or to detect hydrothermal alteration, which in active regions are naturally linked to geothermal activity.

3. METHODOLOGY

3.1 Study area and collected data

Portugal is a Southwestern European country located on the Iberian Peninsula. It borders Spain to the east and north, and the Atlantic Ocean to the west and south.

Mainland Portugal represents the continental part of the country, which has an area of 89 015 km² and does not include the Atlantic islands of Madeira and Azores.

First, the target data (labels) for the machine learning analysis were collected, which consisted of 149 points gathered from the Geothermal Atlas of Portugal. The training labels were provided by LNEG. The data represent different types of wells, such as oil and gas, water, geothermal, geotechnical and mining. They have been drilled inside the territory of Mainland Portugal and also in its continental shelf. At these locations, temperatures were measured at different depths and the geothermal gradient has been calculated at each thermic stable well, avoiding measurements where thermal convection could be happening so the values could be accurate. The geothermal gradient is measured by dividing the difference of the temperatures at different depths by the difference of their depths and is expressed in °C/km. The surface heat flow density was calculated at wells with more than 100m depth that contained more than six different measurements and where the standard deviation values were less than 45%. To calculate it, the thermal conductivity values of each lithological unit were multiplied by the geothermal gradient. The unit is expressed in mW/m².

Only wells inside the territory of Mainland Portugal were used in this work, meaning that the ones located at the continental shelf were excluded. The total amount of points inside the study area was then 138 and their spatial distribution is shown in Figure 1.

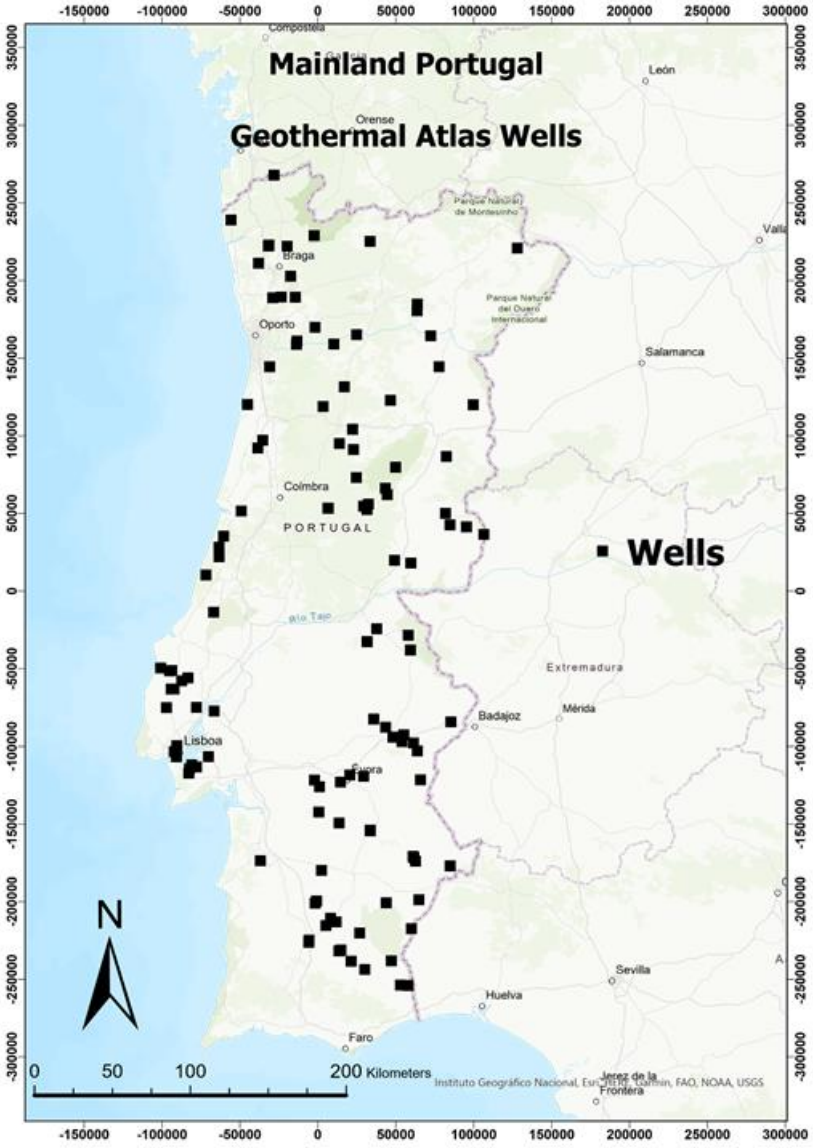

Figure 1. Map showing the study area (Mainland Portugal) and the points locations representing the selected wells from the Geothermal Atlas of Portugal (Ramalho & Correia, 2015).

Secondly, different geospatial data potentially correlated with geothermal activity, as discussed in the literature review, have been gathered to extract important features for the machine learning modelling and to create training data. A summary of the collected dataset is illustrated in Table 1. It includes different types of geological information such as lithology, faults and quaternary faults, soil parental material, soil texture, potential permeability, hydrogeological units, seismic intensity zones, elevation, air temperature, and Bouguer gravity anomaly.

Collected Data	Source	Format
Geothermal wells	LNEG, Geothermal Atlas of Portugal	Shapefile
Elevation (DEM)	SRTM	Raster
Bouguer Gravity Anomaly	IRENA	Raster
Air Temperature 2m	Global Solar Atlas	Raster
Geological Map	LNEG	Shapefile
Faults Map	LNEG	Shapefile
Quaternary Faults	Quaternary Faults Database of Iberia	Shapefile
Soil Parental Material	University of Lisbon`s Superior Institute of Agronomy (ISA)	Shapefile
Soil Texture	University of Lisbon`s Superior Institute of Agronomy (ISA) EPIC WEBGIS	Raster
Potential Permeability	University of Lisbon`s Superior Institute of Agronomy (ISA) EPIC WEBGIS	Raster
Hydrogeological Units	National System of Environment (SNIAmb).	Shapefile
Seimic Intensity Zones	National System of Environment (SNIAmb).	Shapefile

Table 1. All collected data used in this study. The geothermal wells provided the labels, and the remaining data were used to create the features.

The geological map was produced by the Portuguese National Laboratory of Energy and Geology (LNEG), mapped at a 1:500000 scale size, and acquired from LNEG WFS server The geological data has two detail levels of lithological descriptions, as well as geological time (era and eon), system and zone. Different types of rock have distinct thermal properties, which affects the variables of interest, especially the surface heat flow density. The tectonic setting is also an important geological information regarding the thermal capacity, given the fact that tectonic activity is linked to higher geothermal activity. A shapefile containing all types of geological faults as polylines was also collected, as well as another shapefile exhibiting quaternary faults, that represents the most recent geological structures. This type of faults can be correlated to current geological activity. It was collected from the Quaternary Faults Database of Iberia (QAFI).

The used digital elevation model (DEM) is derived from SRTM mission and has cells with 25m resolution. The temperature data comes from the Global Solar Atlas (GSA) and was developed by SOLARGIS. It is a global temperature map (°C) at 2m above the ground. In addition, Bouguer anomaly data was collected from the International Renewable Energy Agency (IRENA). This measurement represents a gravity anomaly that gives information about the Earth's crust thickness. This data is global, derived from the GOCE satellite, part of the Earth Explorer family of earth observation satellites of the European Space Agency (ESA) with a 25km resolution. When the Earth's crust is thinner the geothermal gradient is higher. The unit is measured in m Gal.

The soil composition layer corresponds to a map of the parental material of the soil across Portugal. It was derived from the geological map (1:500000) and it has classifications based on the content of K, Ca, SiO2 and the geological unit texture, containing seven classes. It was acquired from the University of Lisbon's Superior Institute of Agronomy (ISA) portal. Soil texture and Potential Permeability were collected from ISA, at the EPIC WEBGIS Portugal data portal. The information about seismic intensity zones and hydrogeological units were acquired from the National System of Environment (SNIAmb).

3.2 Data preparation

In order to aggregate all data and create tables with features and labels for the machine learning models, vector grids represented by squares of 5x5 and 2,5x2,5 km spatial resolution were built over the study area. The 5x5 grid contains 3830 pixels while the 2,5x,2,5 includes 14810. The grid polygons were created by resampling the 25m Digital elevation model (DEM) of Portugal, which is part of the feature dataset of this study.

The features and labels for this project were built using ArcGIS Pro software, as well as the vector grids with two different spatial resolutions. Most of the preprocessing steps were performed using the same platform while the rest using Python programming language. The grids with all aggregated data were used to create tables, which were then exported to Python environment for the ML modelling.

All raster and shapefiles were projected into the ETRS 1989 Portugal TM06 coordinate system and clipped to cover Mainland Portugal. After resampling the 25m DEM into 2,5x2,5 and 5x5 km, the vector grids were generated using the centroids of their resampled cells. The latitude and longitude coordinates of the centroids, and the value of the resampled elevation were added into the grid as features. This step was performed to provide spatial information into the models.

The geothermal well points were summarized within the vector grids pixels calculating the average of the points values (geothermal gradient and heat flow density) in a pixel. These two variables are the labels, in other words, the variables to be predicted in the analysis. The summary statistics of the labels in each grid are shown in Table 2. The rows containing these values in the grid are the training/test data, and the whole grids were used to generate the model's predictions for all study area. In the 5x5 grid, the

number of rows containing label values is reduced from 138 into 109, while in the 2,5x2,5, to 117. Outliers were removed from the modelling by using the 3 sigma rule (Pukelsheim, 1994). It works by selecting only the data within three times the standard deviation from the mean. Table 3 shows the summary statistics of the labels after outlier removal.

	Geothermal Gradient (°C/km)		Surface Heat Flow Density (mW/m²)	
Grids	2.5x2.5 km	5x5 km	2.5x2.5 km	5x5 km
Count	117	109	117	109
mean	21,976733	22,002821	63,851566	63,419905
std	6,363632	6,119941	22,672707	21,445544
min	10	10,4482	25	25
25%	17,38	17,383486	50,55	50,55
50%	21	21	60,8422	60,99
75%	26	24,7	73,5798	72,5
max	48	48	183,84	183,84

Table 2. Summary statistics of the labels after being aggregated into the vector grids.

	Geothermal Gradient (°C/km)		Surface Heat Flow Density (mW/m²)	
Grids	2.5x2.5 km	5x5 km	2.5x2.5 km	5x5 km
Count	114	107	115	108
mean	21,405946	21,572968	62,150376	62,304904
std	5,338412	5,274973	18,538075	18,095066
min	10	10,4482	25	25
25%	17,245	17,381743	50,3329	50,44145
50%	20,8	20,6	60	60,9161
75%	24,46995	24,5125	73,306	72,30425
max	37	37	114,863	116,65

Table 3. Summary statistics of the labels after removing outliers.

The DEM has also been derived to create other data to be used as features. A drainage map of Portugal was built on ArcGIS Pro and later used to create a drainage density map through Kernel density estimation. A slope map was also generated from the DEM and later reclassified into 16 classes in intervals of 5 degrees of slope, from 0-5° to 75-80°. The faults maps were also used to create Kernel density maps, as well as distance-to maps by using Euclidean distance. A summary of these steps is shown in Figure 2.

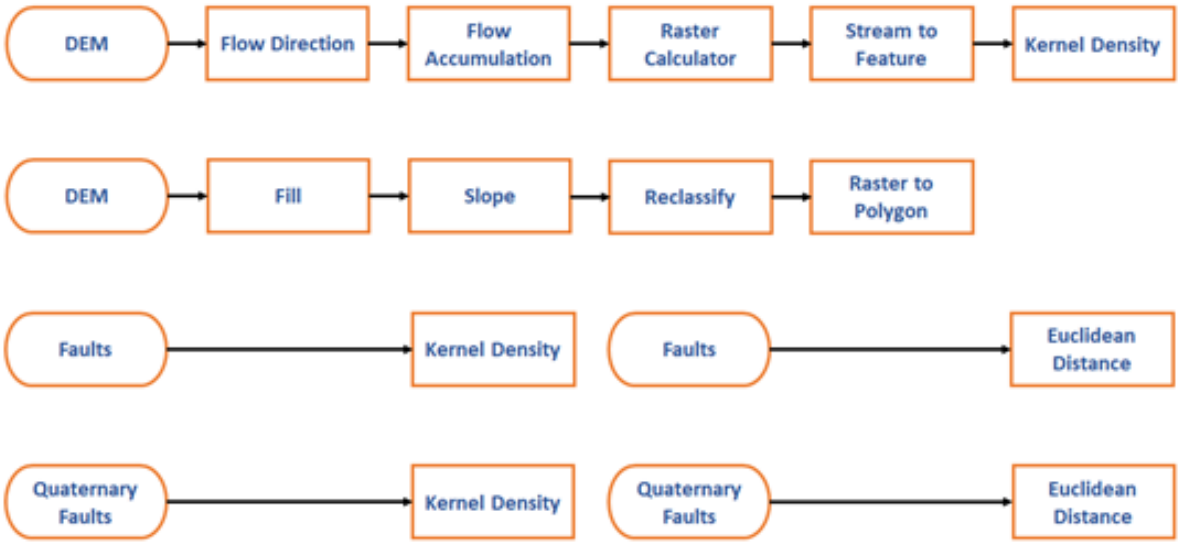

Figure 2. Workflow used to create slope feature, drainage, faults and quaternary faults density, and distance to faults and quaternary faults.

In raster files containing continuous variables, zonal statistics was applied and used to create features like mean, standard deviation, min, max and range. Some statistical features were excluded based on their numeric relevance, as it was the case with the Bouguer anomaly, where only the mean was kept. Since it was resampled from a very low spatial resolution compared to the grids, most statistics did not add any information.

After aggregating the statistics values into the grids, null values were filled in some features using their 8 nearest neighbors mean. This was required to fit the data into the models, since they do not handle null values. For all continuous data, the mean and standard deviation of the eight nearest neighbors were averaged to create new features representing the values of the vicinity, in order to provide more spatial information and improve the accuracy of the ML models.

To prepare categorical values for regression models, that require numeric values in all their features, the one-hot encoding method was performed. This method creates new columns corresponding to each class inside the source column and assign values from 0 to 1 so that 1 is given whenever that class is present and 0 when it is not. When there were null values in the source column, they also became another encoded feature.

The steps to aggregate the information from the shapefiles and categorical raster consisted in intersecting the features with the grids, later dissolved using the index column, and then assigned the maximum value. This way all classes overlapping each

pixel received the number 1, stating their presence in the area. The features columns were then finally joined to the grids. Before performing these steps in categorical raster files, it was necessary to convert them into shapefiles.

A graph summarizing the workflow is shown in Figure 3. In the end, we created 242 features. Most of them are encoded features since every class becomes a column.

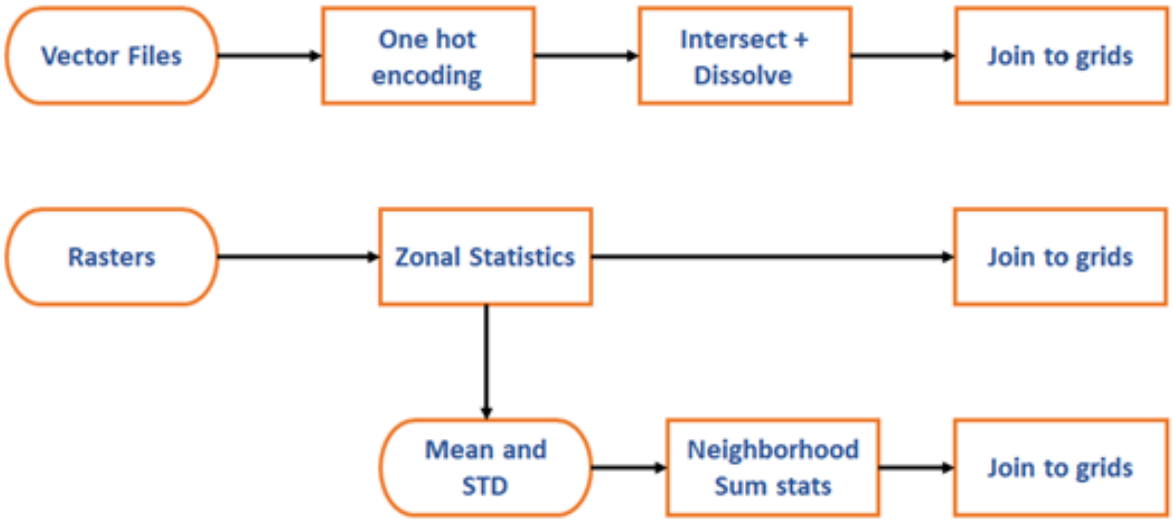

Figure 3. Workflow to aggregate all features into the grids. Categorical rasters were transformed into vectors. Statistical values of the mean and standard deviation of the continuous rasters were averaged among the vicinity and added as features.

3.3 Machine learning models

Two algorithms have been applied to predict the geothermal gradient and the heat flow density throughout Mainland Portugal. Their accuracy and errors were later compared and finally discussed.

Random Forests and Extreme Gradient Boosting (XGB) are both decision-trees algorithms that can be used for classification or regression problems (Breiman, 2001; Chen & Guestrin, 2016). Ensemble machine learning models are characterized by multiple base models that are used to predict a final value. They are known for producing smaller generalization errors. Random Forests utilizes different decision trees running in parallel and returns an estimation of all their outcomes. XGB works by fitting base learners (trees) in a sequenced way into the model. Using multiple trees instead of a single one is less likely to generate overfitting, which happens when a

model performs very accurately for the training data but performs poorly when dealing with unseen data.

Since labels being estimated in this project are both continuous variables, it is a regression modelling problem. The Scikit-learn Python module (Pedregosa et al., 2011) is a machine learning library that contains many state-of-the-art techniques and algorithms made to simplify machine learning and is widely used in scientific papers and research. “Random Forest Regressor” function was used to train the models. Xgboost is another library that provides the model function to be used in many programming languages, including Python. It allows to be used together with Scikit-Learn tools.

3.4 Modelling and evaluation

Using Scikit-Learn tools, models were built, and their accuracy assessed through 5 fold cross-validation (CV) method. Cross-validation splits the data in k number of folds, producing k different training and test datasets. This is used to come up with a more generalizable model, avoiding overfitting, since it explores the dataset in more than only one training/test split. Before applying cross-validation, the dataset was shuffled in Python, in order to remove the vector grids spatial index, therefore creating better spatially distributed training and test datasets. A random state was used assuring the reproducibility of the step. Since the data was previously shuffled in the CV, extra shuffling was not necessary.

Cross-validation was also used to finetune the models hyperparameters. Grid Search CV is a tool where it is possible to create a list of values for the desired hyperparameters and pass it into a cross-validation to select the ones that returns the best average of results. The explored hyperparameters for the Random Forests Regressor were the following: N estimators, random state, max depth, min samples split, min samples leaf, and criterion. For XGB Regressor the hyperparameters were the following: n estimators, random state, learning rate, gamma, reg lambda, and max depth.

The measurements utilized to evaluate the models’ performances were R^2, RMSE, and MAE. MAE is a linear score that stands for mean absolute error and it represents the averaged difference between predicted values and actual values, therefore it is given in the same unit as the predicted variable. It can be significantly increased in the presence

of outliers. RMSE is the root mean squared error and this measurement also represents an averaged distance between predictions and true values, however in terms of the square root of the variance of the residuals. Lower RMSE values in test data represents less overfitted models. R^2 is a coefficient of determination that usually ranges from 0 to 1 and can be interpretated as how far the data are to the fitted regression line. When the value is almost zero it means that the model is not able to determine the variability, while close to one means that the model is able to explain the variability. R^2 compares the fit of a model with a horizontal straight line, to assess the null hypothesis. When the model fits worse than a horizontal line, the R^2 will have a negative value meaning that the model does not follow the data trend. R^2 is used to assess if the model and the output follow a linear correlation.

Scikit-Learn and ensemble algorithms like RF also provide tools for feature selection. After running the models and performing grid search CV, the same process was repeated after removing features based on their importance's to the models, extracted with "Feature Importance's" and "Select from Model" functions. This was done as an attempt of performing dimensionality reduction and assess if the accuracy would improve. Importance is calculated as the normalized reduction of the criterion brought by that feature. Permutation importance was also investigated in order to assess and interpretate the most important features to the models. It measures the importance of features by calculating if the prediction error increases or not errors after permuting each feature. If permuting increases the model error, it means that the feature was relevant to the model. The calculation is done by the distance between a permutation metric and a baseline metric.

After all models were finished, the whole grids were fitted into them to create predictions for the study area. It was later exported as a table to ArcGIS Pro and joined into the grids shapefiles to produce the predicted geothermal gradient and heat flow density maps of Portugal.

4. RESULTS

The complete dataset containing all kept features was shuffled and fitted into a Random Forest Regressor and XGB Regressor for both grids and to predict both labels

(geothermal gradient and heat flow density). Afterwards, a Grid Search Cross-validation was performed to define the best hyperparameters. The accuracy of the models were measured using a 5-fold Cross-validation, that returned the average MAE, RMSE and R^2. Later, the features' importance were used through Select From Model function reducing the number of features. The new models were then tested, and the steps repeated (Grid Search CV and 5-fold CV). The results are shown in the following sections.

4.1 Random Forests models

Final hyperparameters and errors measurements for Random Forests models are summarized in Table 4 (geothermal gradient) and Table 5 (heat flow). The best performance when predicting geothermal gradient was achieved with the 5x5 km spatial resolution grid using the features selected based on their importance in the first model. This model reduced the number of features from 242 to 40. All RF model's prediction maps covering the study area are shown in Figure 4 (geothermal gradient).

Geothermal Gradient				
	Random Forests (5x5 grid)		Random Forests (2.5x2.5 grid)	
Hyperparameters	All Features	Selected from Model	All Features	Selected from Model
Random State	3	2	2	2
Criterion	Squared error	Squared error	Absolute error	Absolute error
N Estimators	100	250	100	100
Max Depth	5	5	5	5
Min Samples Leaf	1	2	1	2
Min Samples Split	3	2	2	2
Errors				
MAE	3,746717631	3,66887079	4,058596306	4,016927566
RMSE	4,7220471	4,706783957	5,151928474	5,137863459
R2	0,201384987	0,215191652	0,039130307	0,045124678

Table 4. Hyperparameters and error measurements for every RF model when predicting geothermal gradient. Hyperparameters were obtained with Grid search CV and errors obtained with K fold CV (k=5). Models used all features and features selected based on their importance scores in the models.

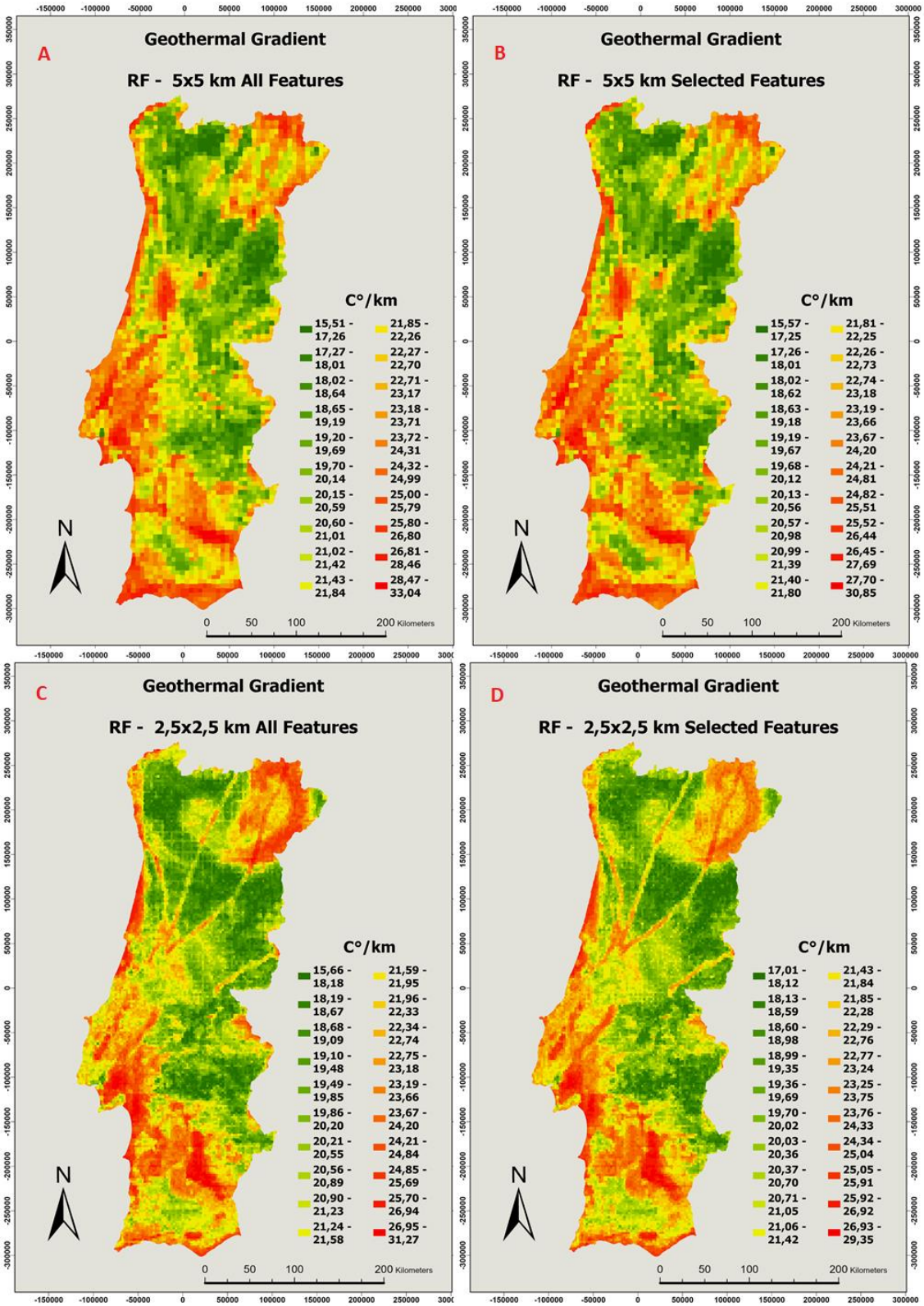

Figure 4. Geothermal gradient prediction maps with Random Forests. **A.** Model using 5x5 km grid and all features. **B.** Model using 5x5 km grid and selected features. **C.** Model using 2,5x2,5 km grid and all features. **D.** Model using 2,5x2,5 km grid and selected features.

When predicting surface heat flow density, the best performance was also achieved with the 5x5 km spatial resolution grid and using features selected based on their importance in the first model. The number of features reduced to 59. All RF model's prediction maps covering the study area are shown in Figure 5 (heat flow density).

Heat Flow Density				
	Random Forests (5x5 grid)		Random Forests (2.5x2.5 grid)	
Hyperparameters	All Features	Selected from Model	All Features	Selected from Model
Random State	4	2	4	3
Criterion	Absolute error	Absolute error	Absolute error	Squared error
N Estimators	250	200	100	100
Max Depth	10	15	15	10
Min Samples Leaf	1	1	2	1
Min Samples Split	2	3	2	3
Errors				
MAE	12,38395917	12,0677001	13,06006468	12,92838859
RMSE	15,33868577	15,03292005	16,13293413	16,09533897
R2	0,221323496	0,248328421	0,223639738	0,231819702

Table 5. Hyperparameters and error measurements for every RF model when predicting heat flow density. Hyperparameters were obtained with Grid search CV and errors obtained with K fold CV (k=5). Models used all features and also features selected based on their importance scores in the models

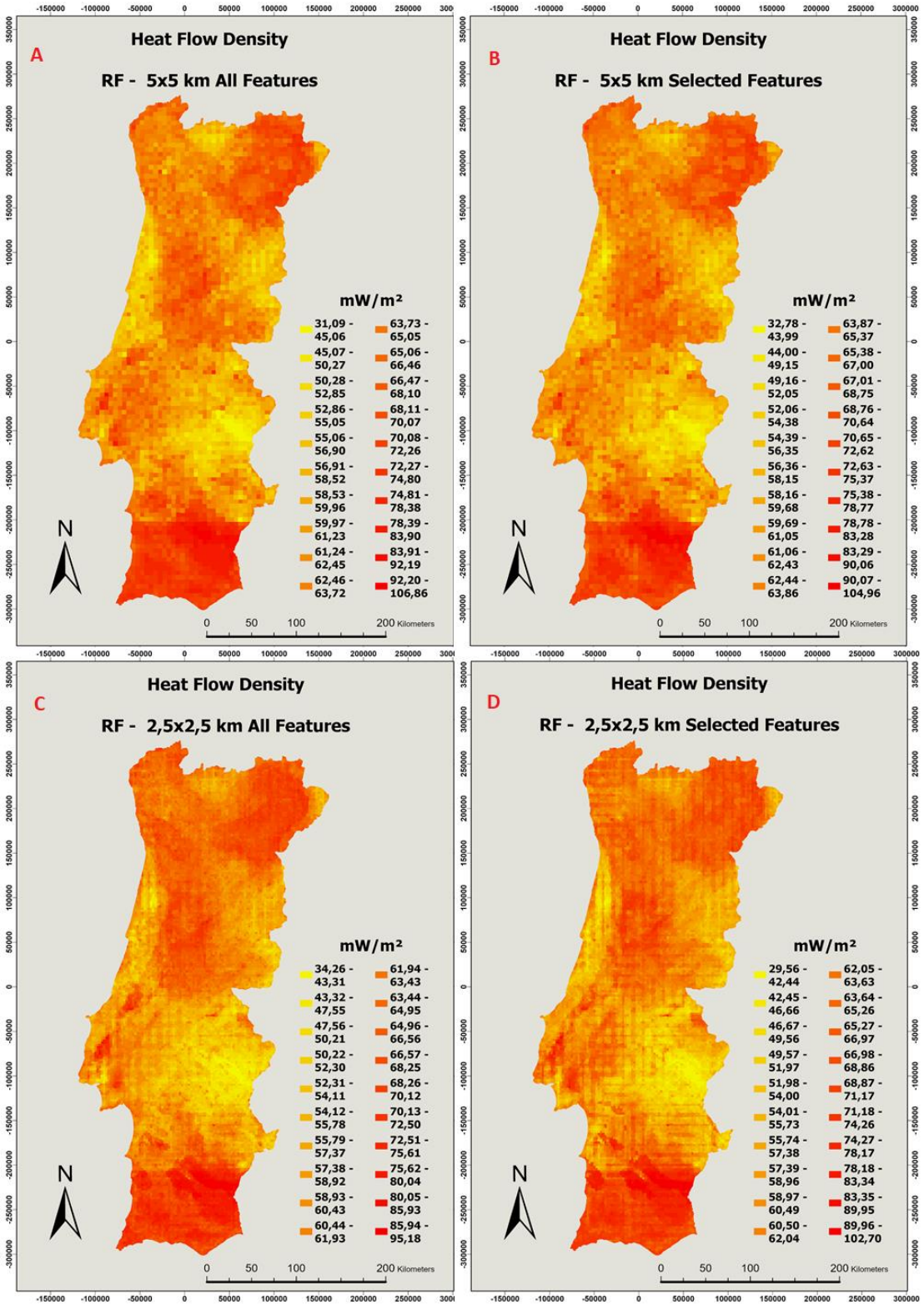

Figure 5. Heat flow density at surface prediction maps with Random Forests. **A.** Model using 5x5 km grid and all features. **B.** Model using 5x5 km grid and selected features. **C.** Model using 2,5x2,5 km grid and all features. **D.** Model using 2,5x2,5 km grid and selected features.

4.2 XGB models

The same methodology was applied with XGB Regressor. Table 6 and 7 show the results for the models when predicting geothermal gradient and heat flow, respectively. Best performance predicting geothermal gradient was also achieved in a 5x5 km grid and with feature selection. This reduced the number of features to 45. When predicting geothermal gradient with XGB, the models using data aggregated into a 2,5x2,5 km grid showed negative R^2 values and considerably higher MAE and RMSE. For this reason, they were not used to create prediction maps. The maps are shown in Figures 7 (geothermal gradient). Best performance predicting heat flow density was also achieved with 5x5 km grid fitted with selected features. The number of features reduced to 46. The predictions maps of heat flow density are shown in Figure 8.

Geothermal Gradient				
	XGB (5x5 grid)		XGB (2.5x2.5 grid)	
Hyperparameters	All Features	Selected from Model	All Features	Selected from Model
Random State	2	2	2	2
Learning Rate	0,1	0,05	0,01	0,05
N Estimators	500	100	500	100
Max Depth	15	12	5	5
Gamma	10	10	10	10
Reg Lambda	10	1	0,1	0,1
Errors				
MAE	3,787958992	3,614766839	4,356813204	4,275029737
RMSE	4,812866168	4,778081377	5,420207974	5,340590027
R2	0,14861505	0,204343513	-0,069126467	-0,043858095

Table 6. Hyperparameters and error measurements for every XGB model when predicting geothermal gradient. They were obtained with Grid search CV and errors obtained with K fold CV (k=5). Models used all features and also features selected based on their importance scores in the models.

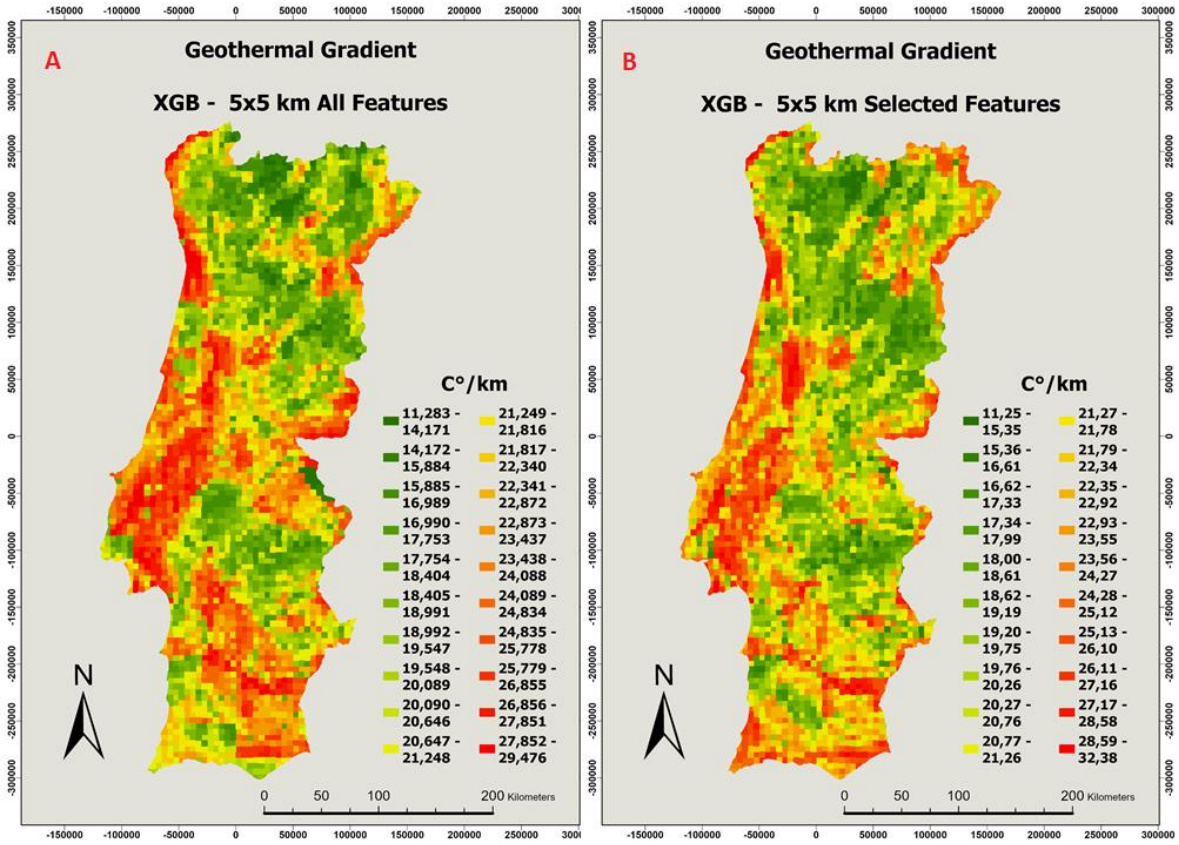

Figure 6. Geothermal gradient prediction maps estimated using 5x5 km grid with XGB. **A.** Model using all features. B. Model using selected features.

Heat Flow Density				
	XGB (5x5 grid)		XGB (2.5x2.5 grid)	
Hyperparameters	All Features	Selected from Model	All Features	Selected from Model
Random State	2	2	2	2
Learning Rate	0,1	0,01	0,05	0,01
N Estimators	200	500	100	500
Max Depth	5	5	5	5
Gamma	10	0,1	0,1	10
Reg Lambda	0,1	0,1	1	1
Errors				
MAE	12,26850403	12,02276277	12,91684395	12,93098171
RMSE	15,71138418	15,30638079	16,39341767	16,63282622
R2	0,170151418	0,220288661	0,209721412	0,194997342

Table 7. Hyperparameters and error measurements for every XGB model when predicting heat flow density. They were obtained with Grid search CV and errors obtained with K fold CVs (k=5). Models used all features and also features selected based on their importance scores in the models.

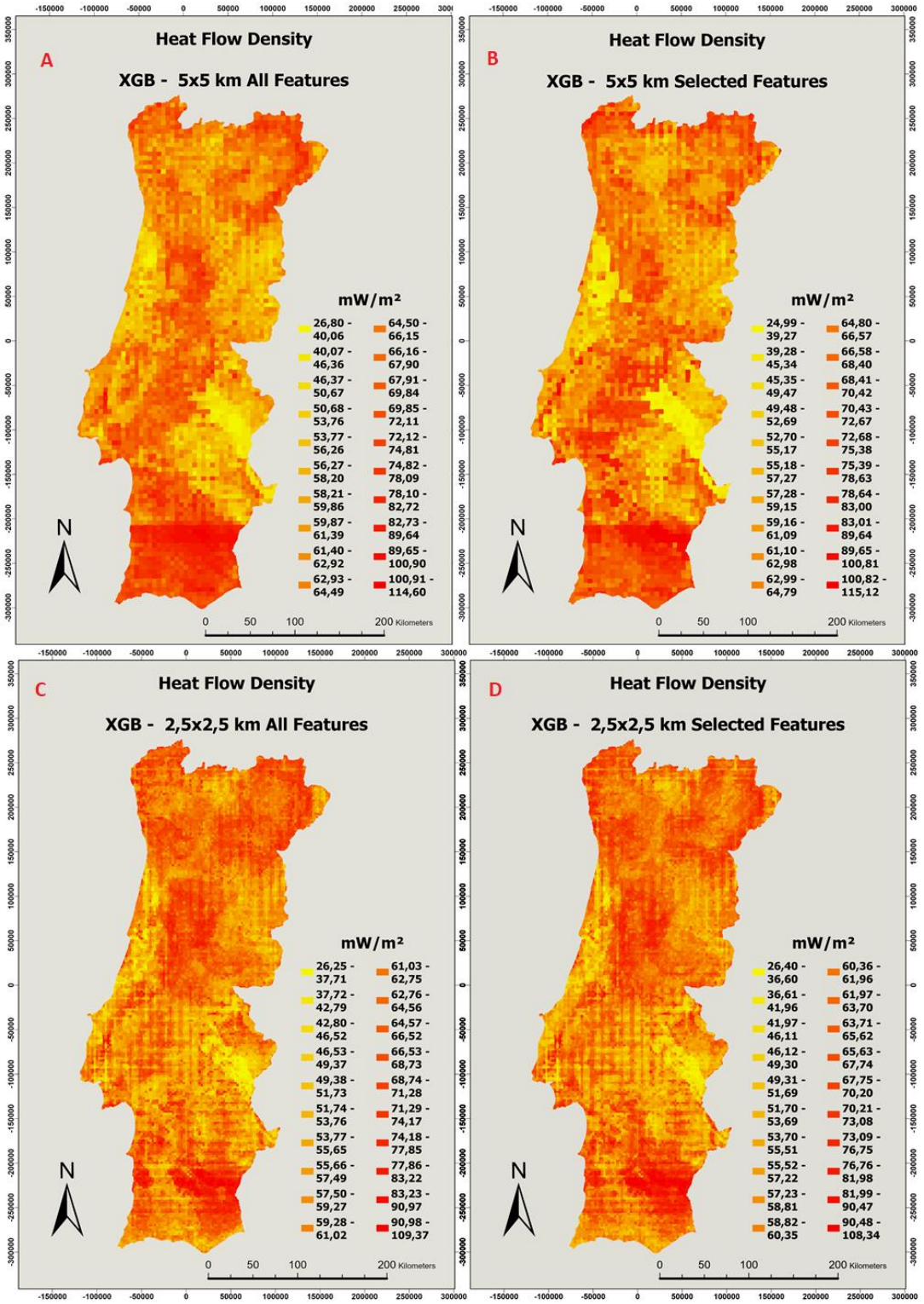

Figure 7. Heat flow density at surface prediction maps with XGB. **A.** Model using 5x5 km grid and all features. **B.** Model using 5x5 km grid and selected features. **C.** Model using 2,5x2,5 km grid and all features. **D.** Model using 2,5x2,5 km grid and selected features.

5. DISCUSSIONS

5.1 Model performance

The results evidenced that Random Forests overall performed slightly better than XGB when predicting both labels. The data aggregation into the lower spatial resolution grid (5x5 km) managed to return better results. This could be related to the presence of data sources with low spatial resolution used as features in the modelling. Table 8 compares the results of the models fitted with selected features in each grids with both algorithms. It is possible to see that RF had slightly lower RMSE values and higher R^2 values compared to XGB. The MAE was lower than XGB's in most models, with few exceptions.

	Geothermal gradient		Heat flow density	
5x5 grid:	RF	XGB	RF	XGB
MAE	3,669	3,615	12,068	12,023
RMSE	4,707	4,778	15,033	15,306
R2	0,215	0,204	0,248	0,220
2,5x2,5 grid:	RF	XGB	RF	XGB
MAE	4,017	4,275	12,928	12,931
RMSE	5,138	5,341	16,095	16,633
R2	0,045	-0,044	0,232	0,195

Table 8. Comparison between Random Forests and Extreme Gradient Boosting results using both grids and selected features.

By analyzing the models' accuracies, we can determine that despite relatively high RMSE values, which can be interpreted that the models were not able to establish really strong correlations to the labels trends and their generalization capacity is not too strong, the models managed to estimate the overall spatial distribution of low and high values of the labels correctly, and the mean absolute error was low enough for detecting potential zones for geothermal surveys. Modelling subsurface properties like geothermal gradient using mostly surface related features seems to be only achievable with some error. The general low values of R^2 indicates that the labels' trend do not follow a strong linear correlation to the data, therefore MAE and RMSE are better metrics to evaluate the models prediction capacity. In this study, the best MAE average between predicted and measured geothermal gradient was 3,61°C/km, with predicted

values ranging from 11,35 to 32,38 (°C/km). As an example, Shahdi et al., (2021) calculated the mean absolute error between predicted geothermal gradient and temperature at specific depths using XGB and Deep Neural Networks and achieved MAE scores of 5,6 and 7 (°C/km), respectively. When evaluating their models with their main dataset to predict subsurface temperatures, they achieved MAE of 3,21 and RMSE of 4,94 with XGB, and MAE of 3,25 and RMSE of 5.01 with RF.

5.2 Features importance and relationships

Based on the features with higher scores of importance and permutation importance, it was possible to identify what seemed to be the most promising features to model geothermal gradient and heat flow density. Figure 11 shows plots of the 20 highest features scores for RF models with data aggregated in a 5x5 km grid. They were calculated using RF feature importance function and permutation importance.

The mean of the standard deviation regarding neighborhood's distance to quaternary faults was given the highest score in both importance calculations when predicting geothermal gradient. Among the most important features, the criteria were: distance to quaternary faults, elevation, density and distance to faults, temperature, Bouguer anomaly and density of drainages.

To predict heat flow density, latitude had the higher importance score in both importance calculations. The most important features were: latitude, distance to quaternary faults, distance and density of faults, density of drainages, soil texture (coarse), soil composition (basic heavy), lithology (Pyrite belt), geochronology systems (Cambrian Ordovician), Bouguer anomaly, longitude and surface air temperature (2 m).

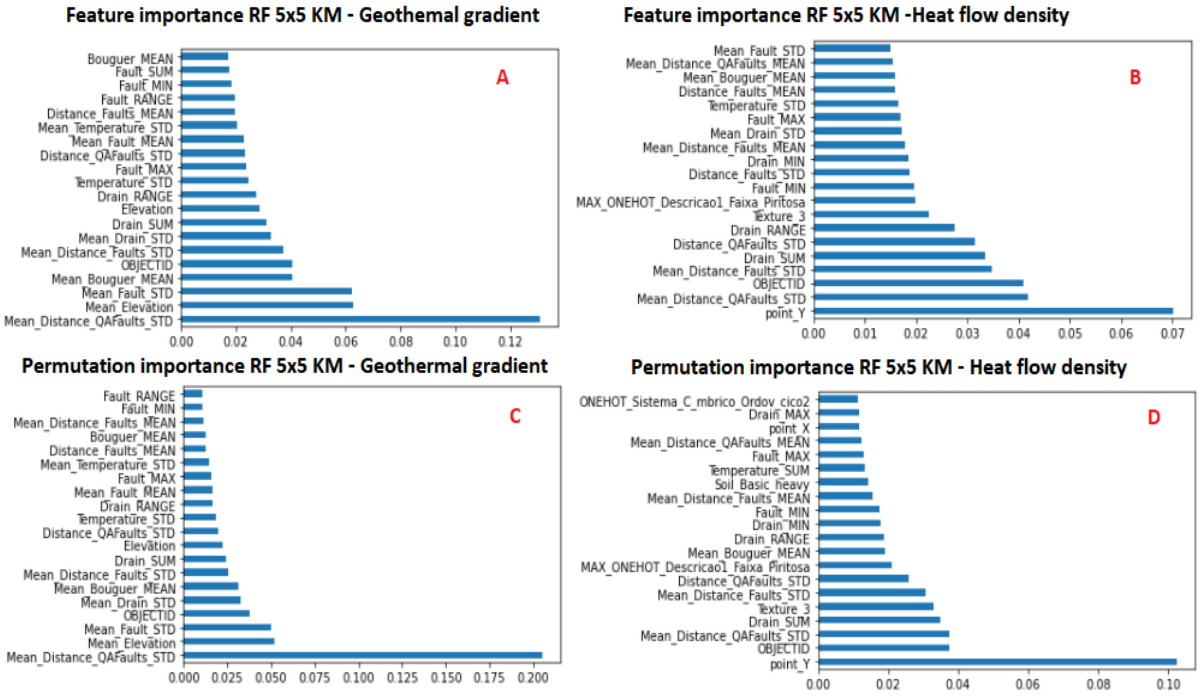

Figure 8. Graphs showing the 20 highest scores for the RF feature importance and permutation importance functions. **A and B.** Feature importance scores for the geothermal gradient and heat flow density, respectively. **C and D.** Permutation importance scores for the geothermal gradient and heat flow density, respectively.

The most important features contributing to the modelling are mainly geological, hydrogeological and geophysical parameters. Surface heat flow density is naturally more influenced by surface properties than geothermal gradient. Soil texture, soil composition, and lithology (due to distinct thermal conductivities) are related parameters. The results of the importance scores seems to demonstrate that as well. The presence of high values of heat flow density in the Pyrite belt in the south of Portugal could be one of the factors that contributed to the great importance given to latitude.

Geothermal gradient, on the other hand, seems to be more related to subsurface parameters. However, they could be linked to the presence of faults, especially recent (quaternary) faults, and also high density of drainages and lower Earth's crust thickness (Bouguer anomaly). In the prediction maps, the influence given to quaternary faults can be seen by higher values close to where there are fault traces (Figure 4).

5.3 Limitations and future work

Although this study used cross validation, creating different folds of training and test data, it did not use an extra dataset with known labels' values to be used for validation of the models. This happened due to the small number of wells compared to the study area. The regions in the study area that do not have training data are likely to have less accuracy. Errors associated with the measurements of the labels at the wells are also present in this study. The measured temperatures at the wells, despite quality assurance methods, are subject to external influences, such as possible water circulation. The estimations of the heat flow depend if the thermal conductivity of the lithological units were measured at the wells or if a theoretical value was assigned based on literature and associated known values.

Improving the accuracy of the models could be achieved with more and better spatially distributed training data, as well as by fitting into the model other important features related to geothermal parameters. Adding more data regarding physical and chemical parameters of the ground will probably improve the model prediction capacity. Remote sensing images such as thermal infra-red, radar and multispectral images should also be explored, since they have many applications in the geothermal field (Booysen et al., 2021; van der Meer et al., 2014). Higher resolution features with more detail could also be inputted and assessed if they are able to return better results.

Exploring different types of algorithms, such as deep learning, is a promising path to be pursed. A comparison with geostatistical methodologies could also be performed in terms of accuracy and spatial distribution. However, geo statistics are not able to create prediction models.

6. CONCLUSIONS

This study utilized two ensemble machine learning algorithms to predict crucial thermal parameters of the surface and subsurface of Mainland Portugal: geothermal gradient and surface heat flow density. Data from 138 wells where the two labels have been measured were aggregated into two vector grids with distinct spatial resolutions (5x5 km and 2,5x2,5 km). Prediction maps covering Mainland Portugal were built using the trained models. They were able to target regions with higher geothermal potential and

can be applied to aid geothermal exploration surveys. The models were also able to point to important features linked to the two variables and confirmed the relevance of geological, hydrogeological and geophysical data, especially most recent fault structures. The comparison between Random Forests and Extreme Gradient Boosting revealed that RF achieved better results overall and the comparison between different spatial resolutions revealed better achievements when aggregating data in a lower scale grid (5x5 km). RF and 5x5 grid had the best MAE of 3,67 °C/km when predicting geothermal gradient and 12,07 mW/m² when predicting surface heat flow density. XGB and 5x5 grid achieved a MAE of 3,61 °C/km when predicting geothermal gradient and 12,02 when predicting the surface heat flow density. However, R^2 values were higher for the RF models compared to the XGB's and in most cases the MAE and RMSE values were lower.

The goal of the study was to explore machine learning methods an assess its applicability in the geothermal field. This study did not intend to determine if ML can be more accurate than other methods, such as physics based or geo statistical models. The results evidenced potential in the methodology, and suggestions on how to improve and future works possibilities were discussed.

Bibliographic References

Ammann, M. W., Walker, A. M., Stackhouse, S., Wookey, J., Forte, A. M., Brodholt, J. P., & Dobson, D. P. (2014). Variation of thermal conductivity and heat flux at the Earth's core mantle boundary. *Earth and Planetary Science Letters*, *390*, 175–185. https://doi.org/10.1016/j.epsl.2014.01.009

Arola, T., Korhonen, K., Martinkauppi, A., Leppäharju, N., Hakala, P., Ahonen, L., & Pashkovskii, M. (2019). Creating shallow geothermal potential maps for Finland using finite element simulations and machine learning. *European Geothermal Congress 2019*, *June*, 6.

Assouline, D., Mohajeri, N., Gudmundsson, A., & Scartezzini, J. L. (2019). A machine learning approach for mapping the very shallow theoretical geothermal potential. *Geothermal Energy*, *7*(1). https://doi.org/10.1186/s40517-019-0135-6

Badino, G. (2005). Underground drainage systems and geothermal flux. *Acta Carsologica*, *34*(2), 277–316. https://doi.org/10.3986/ac.v34i2.261

Barbier, E. (2002). Geothermal energy technology and current status: An overview. *Renewable and Sustainable Energy Reviews*, *6*(1–2), 3–65. https://doi.org/10.1016/S1364-0321(02)00002-3

Bignall, G., Milicich, S., Ramirez, E., Rosenberg, M., Kilgour, G., & Rae, A. (2010). Geology of the Wairakei-Tauhara Geothermal System, New Zealand. *Proceedings World Geothermal Congress*, *November 2019*, 25–29.

Booysen, R., Gloaguen, R., Lorenz, S., Zimmermann, R., & Nex, P. A. M. (2021). Geological Remote Sensing. In *Encyclopedia of Geology* (2nd ed.). Elsevier Inc. https://doi.org/10.1016/b978-0-12-409548-9.12127-x

L. Breiman, "Random forests", Mach. Learn., vol. 45, no.1, pp. 5-32, Oct. 2001. doi:10.1023/A:1010933404324

Chamorro, C. R., García-Cuesta, J. L., Mondéjar, M. E., & Linares, M. M. (2014). An estimation of the enhanced geothermal systems potential for the Iberian Peninsula. *Renewable Energy*, *66*, 1–14. https://doi.org/10.1016/j.renene.2013.11.065

Corniello A., D. D. (2012). *Hydrogeology key role in a geothermal exploration in southern Italy*. *January*. http://www.researchgate.net/publication/255486060_Hydrogeology_key_role_in_a_geothermal_exploration_in_southern_Italy

Coro, G., & Trumpy, E. (2020). Predicting geographical suitability of geothermal power plants. *Journal of Cleaner Production*, *267*, 121874. https://doi.org/10.1016/j.jclepro.2020.121874

Fahil, A. S., Ghoneim, E., Noweir, M. A., & Masoud, A. (2020). Integration of well logging and remote sensing data for detecting potential geothermal sites along the gulf of Suez, Egypt. *Resources*, *9*(9). https://doi.org/10.3390/RESOURCES9090109

Gangwani, P., Soni, J., Upadhyay, H., & Joshi, S. (2020). A Deep Learning Approach for Modeling of Geothermal Energy Prediction. *International Journal of Computer Science and Information Security*, *18*(1), 62–65.

González, D. L., & Rodríguez-Gonzálvez, P. (2019). Detection of geothermal potential zones using remote sensing techniques. *Remote Sensing*, *11*(20). https://doi.org/10.3390/rs11202403

Han, M., Feng, Y., Zhao, X., Sun, C., Hong, F., & Liu, C. (2019). A Convolutional Neural Network Using Surface Data to Predict Subsurface Temperatures in the

Pacific Ocean. *IEEE Access*, *7*, 172816–172829. https://doi.org/10.1109/ACCESS.2019.2955957

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences (Switzerland)*, *10*(17). https://doi.org/10.3390/app10175975

Lerche, I. (1991). Temperature dependence of thermal conductivity and its impact on assessments of heat flux. *Pure and Applied Geophysics PAGEOPH*, *136*(2–3), 191–200. https://doi.org/10.1007/BF00876371

Lund, J. W., & Freeston, D. H. (2001). World-wide direct uses of geothermal energy 2000. *Geothermics*, *30*(1), 29–68. https://doi.org/10.1016/S0375-6505(00)00044-4

Merembayev, T., Yunussov, R., & Yedilkhan, A. (2019). Machine learning algorithms for classification geology data from well logging. *14th International Conference on Electronics Computer and Computation, ICECCO 2018*, 9–13. https://doi.org/10.1109/ICECCO.2018.8634775

Mock, J. E., Tester, J. W., & Wright, P. M. (1997). Geothermal energy from the earth: Its potential impact as an environmentally sustainable resource. *Annual Review of Energy and the Environment*, *22*(1), 305–356. https://doi.org/10.1146/annurev.energy.22.1.305

Mohamed, H. S., Abdel Zaher, M., Senosy, M. M., Saibi, H., el Nouby, M., & Fairhead, J. D. (2015). Correlation of Aerogravity and BHT Data to Develop a Geothermal Gradient Map of the Northern Western Desert of Egypt using an Artificial Neural Network. *Pure and Applied Geophysics*, *172*(6), 1585–1597. https://doi.org/10.1007/s00024-014-0998-1

Nguyen, A., Pasquier, P., & Marcotte, D. (2017). Borehole thermal energy storage systems under the influence of groundwater flow and time-varying surface temperature. *Geothermics*, *66*, 110–118. https://doi.org/10.1016/j.geothermics.2016.11.002

Olasolo, P., Juárez, M. C., Morales, M. P., Damico, S., & Liarte, I. A. (2016). Enhanced geothermal systems (EGS): A review. *Renewable and Sustainable Energy Reviews*, *56*, 133–144. https://doi.org/10.1016/j.rser.2015.11.031

Omwenga, B. M. (2020). Geothermal Well Site Suitability Selection Using Geographic Information Systems (GIS) and Remote Sensing: Case Study of the Eburru Geothermal Field. *45th Workshop on Geothermal Reservoir Engineering*, *1*, 1–6.

F. Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825–2830, 2011.

Pukelsheim. (1994). The three sigma rule. *American Statistician*, *48*(2), 88–91. https://doi.org/10.1080/00031305.1994.10476030

Ramalho, E. C. (2014). O papel do atlas geotérmico nacional no fomento da exploração da energia geotérmica em Portugal continental. *Comunicacoes Geologicas*, *101*, 833–836.

Ramalho, E., Correia, A., 2015. Atlas Geotérmico de Portugal Continental [mapa] https://geoportal.lneg.pt/mapa/?mapa=AtlasGeotermico [20 Fev, 2022].

Rezvanbehbahani, S., Stearns, L. A., Kadivar, A., Walker, J. D., & van der Veen, C. J. (2017). Predicting the Geothermal Heat Flux in Greenland: A Machine Learning Approach. *Geophysical Research Letters*, *44*(24), 12,271-12,279. https://doi.org/10.1002/2017GL075661

Robertson, E. C. (1988). Thermal Properties of Rocks. *US Department of the Interior: Geological Survey*, 88–441.

S. Angra and S. Ahuja, "Machine learning and its applications: A review," in 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE, 2017, pp. 57–60.

Shahab, A., & Singh, M. P. (2019). Comparative analysis of different machine learning algorithms in classification of suitability of renewable energy resource. *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019*, *2013*, 360–364. https://doi.org/10.1109/ICCSP.2019.8697969

Shahdi, A., Lee, S., Karpatne, A., & Nojabaei, B. (2021). Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. *Geothermal Energy*, *9*(1). https://doi.org/10.1186/s40517-021-00200-4

Sun, Z., Jiang, B., Li, X., Li, J., & Xiao, K. (2020). A data-driven approach for lithology identification based on parameter-optimized ensemble learning. *Energies*, *13*(15), 1–15. https://doi.org/10.3390/en13153903

Suwai, J. (2009). Basic Hydrogeology in Geothermal Systems. *Short Course V on Exploration for Geothermal Resources*, 1–9.

Tester, J. (2007). *Congressional Testimony for*. 1–13.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785–794.

Tinti, F., Kasmaee, S., Elkarmoty, M., Bonduà, S., & Bortolotti, V. (2018). Suitability evaluation of specific shallow geothermal technologies using a GIS-Based multi criteria decision analysis implementing the analytic hierarchic process. *Energies*, *11*(2). https://doi.org/10.3390/en11020457

Trumpy, E., Donato, A., Gianelli, G., Gola, G., Minissale, A., Montanari, D., Santilano, A., & Manzella, A. (2015). Data integration and favourability maps for exploring geothermal systems in Sicily, southern Italy. *Geothermics*, *56*(January 2021), 1–16. https://doi.org/10.1016/j.geothermics.2015.03.004

van der Meer, F., Hecker, C., van Ruitenbeek, F., van der Werff, H., de Wijkerslooth, C., & Wechsler, C. (2014). Geologic remote sensing for geothermal exploration: A review. *International Journal of Applied Earth Observation and Geoinformation*, *33*(1), 255–269. https://doi.org/10.1016/j.jag.2014.05.007

Zhang, L., & Zhan, C. (2017). *Machine Learning in Rock Facies Classification: An Application of XGBoost*. *January*, 1371–1374. https://doi.org/10.1190/igc2017-351

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS