Erasmus University Rotterdam

Erasmus School of Economics

MSc Quantitative Marketing and Business Analytics

Master Thesis

# Epidemic Risk Modelling: Dengue Fever in the Philippines

*Author:*
Elena-Andreea Stan

*Student Number:*
427743

*Supervisor:*
Dr. Philip Hans Franses

*Second assessor:*
Dr. Kathrin Gruber

*Date:* August 6, 2019

# Contents

# 1 Introduction

Epidemics of infectious diseases are one of the most destructive and costly, globally occurring, hazardous phenomena. The current response to the spread of epidemics is mainly reaction focused. This approach has proven to be not only slow but also expensive and outdated. Taking the spread of the Ebola virus in Western Africa as a lesson, the World Health Organization learned that better preparation and the putting in place of sensitive epidemic surveillance is the first step towards better control of outbreaks of infectious diseases in the less developed parts of the world (Kekulé, 2015). Organizations such as the Netherlands Red Cross have begun investing in an Epidemic Risk Assessment project (ERA) in order to increase prevention and preparedness (The Netherlands Red Cross, 2018). Models that fit the specific and sensitive nature of the data of affected locations have become a necessity. Working with humanitarian data sets, however, poses new challenges for statistical analysis.

Presently the study into epidemic risk has run into insufficient availability of data due to time constraints or simply missing values for the desired granularity. Therefore, calculating weights for epidemic risk parameters has proven to be an intricate complication. The currently available research on infectious diseases, more specifically into environmental risk factors of epidemics, has the major disadvantage that it lacks interpretability. For example, during a study on Dengue incidence in the Philippines missing data on provinces affected the normalization step where using less regions led to an erroneous distribution of risk (Hierink, 2018).

This paper proposes a study into these issues. The objective is to model epidemic risk by introducing methods of handling missing data patterns. By addressing each particular missingness mosaic with an isolated procedure for imputation such a goal is achieved. Previous epidemiological studies choose Complete Case Analysis (CCA) to work around such problems (Liu and De, 2015) or opted for one imputation method throughout the entire course of the study: k-Nearest-Neighbours (kNN) in Hierink (2018) or Model Based Imputation in Harel et al. (2017). This paper introduces a new method, namely one that handles each pattern of missingness accordingly. Furthermore, it compares the new Pattern Based Imputation with the method used primarily by The Netherlands Red Cross ERA: full kNN Imputation. This study will prove the superiority of the mosaic-based imputation over the full kNN. In order to do so, the types of variables must first be defined.

The European Commission developed a composite indicator of risk, INFORM, which serves as a measure of the global distribution of risk of crises and disasters (De Groeve et al., 2016). This indicator is built out of three dimensions: Hazard & Exposure, Vulnerability and Lack of Coping Capacity. Each one of these levels contains a number of categories which encompass predefined types of variables. Present research found three main structures amongst the INFORM components: Independent Missing Variables (IMV), Consequential Missing Variables (CMV) and Regular Missing Variables (RMV). IMV's include naturally occurring variables such as weather, tsunamis, floods, cyclones etc. all found within the Hazard & Exposure dimension. CMV's represent variables which come as a direct consequence of IMV's: development deprivation, aid dependency etc. These types can be found under the Vulnerability dimension. Lastly, RMV's represent directly human influenced variables found within the Lack of Coping Capacity dimension. Examples of such variables include governance, communication, infrastructure and access to health care system.

The present research plans on using Red Cross provided data sets on monthly new Dengue fever cases and influential factors in the Philippines. Furthermore it tries to improve the reliability of the study results which contain the difficult variables. The before mentioned ERA framework has categorized a wide range of risk factors as well as done preliminary research into their influence. It is believed that issues such as underreporting could be one of the causes of the missing data points, hence one can categorize the unavailable observations as Missing Not At Random (MNAR). This type of missingness describes the absence of data points which depend on some unobserved predictor, in the present case underreporting. Dealing with this type of unavailability is particularly tricky as imputation introduces bias in inference models (Gelman and Hill, 2006).

Therefore the main aim of the research present in this paper is developing expert based weights for all indicators of infectious diseases in difficult data sets. Obstacles that have been overcome include: high percentages of missing data, overdispersion and multicollinearity. This study first determines what is the optimal way to find and handle missing values in each of the three types of variables. Important considerations such as how the patterns can be identified, which imputation method best fits each mosaic and whether the statistical properties are preserved will be examined and appropriate advice will be provided. Furthermore, three modelling techniques are put forward: a Selection Model which uses Elastic Net regularization, a newly developed Explanatory Model for

the selected variables taking into account over-dispersion, namely the Quasi-Poisson Regression on Elastic Net Selected Variables and a Predictive Model via Bayesian Neural Networks. The high uncertainty in the data due to the imputation as well as the multicollinearity issues that arise from the types of variables in epidemic studies lead to the fact that the usual statistical techniques fail to select, explain and predict accordingly. The new Explanatory technique will prove to outperform the Selection Model in terms of model fit and the Predictive Model in terms of forecasting. What this study adds to current literature is first the introduction of a new approach to variable imputation. Second, never before attempted analysis is performed using model selection via Elastic Net regularization and a subsequent Quasi-Poisson regression on the selected variables. Thus, viable coefficients as well as correct standard errors are obtained. Last but not least, this study will introduce the first Bayesian Neural Network analysis within the field of epidemic studies.

This paper is structured as follows: first a Data section dives into a description of the data set as well as the difficulties it brings, then a Literature review will present what is the background of the current research and what is being added to it through this study. How the data will be handled can be found in the Methodology section and, lastly, a Conclusion will be summing up the main findings and give suggestions for further research.

## 2 Data

The data set consists of monthly data on 40 continuous variables resulting in a total of 960 observations describing 80 regions of the Philippines. These variables were recorded per region. Having as dependent variable the amount of new cases of Dengue fever, the independent variables describe features such as population number according to the respective gender, weather, farm animals, type of water sewer tank, indicator for open/closed pit, water source and population density. The data is recorded in the year 2015. Figure 1 displays the dependent variable per province. A detailed description of each variable can be found in the Appendix Table 4.

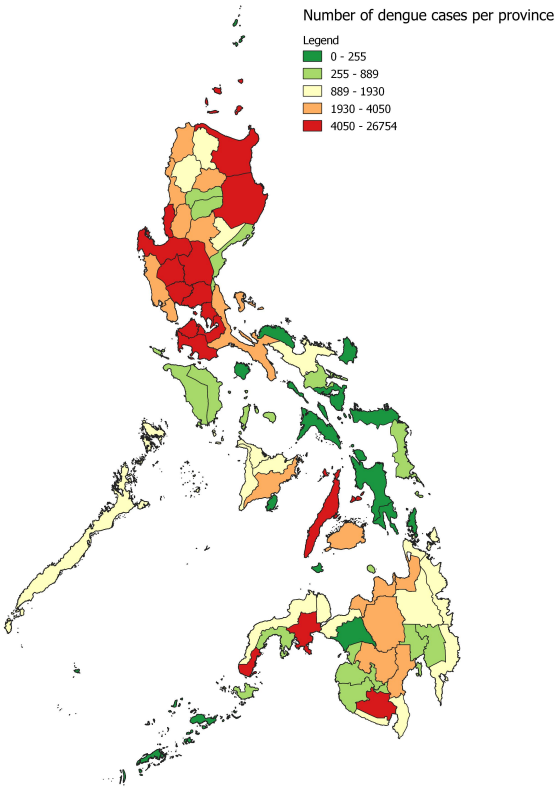



Figure 1: Total Dengue cases per province in 2015.

All features are categorized within the Epidemic Risk Assessment (ERA). They represent drivers for Infectious Hazards & Exposure, Vulnerability and Lack of Coping Capacity within the INFORM index (De Groeve et al., 2016). Variables within the data set are particularly challenging as much information is missing. The first step in understanding the missing values within the data set contains determining the amount of non-available observation points present. Table 4 in the Appendix

showcases the percentage of missing values per variable. It can be seen that groups of variables seem to have very close percentage of missingness. This could indicate an underlying pattern in the mechanism of missing values within the data set. There seem to be weather variables with an average of 45% missing value, soil related features with an average of 63% missingness, farm animals with 2.5% non-available information, toilet & water availability traits with 12.5% missing values and the population density measure with 2.5% *NA* values. In order to gain better insight into what causes these percentages to be classifiable into categories based on the percentage of missingness, visualization techniques are necessary. Templ et al. (2012) put forward a plot which showcases all combinations of (non-)missing values in the data set. This is called an 'aggregation plot' and is graphed in Figure 2. When two variables contain unavailable information for the same observation, their combination is signalled out through a different colour, in the present case dark grey.

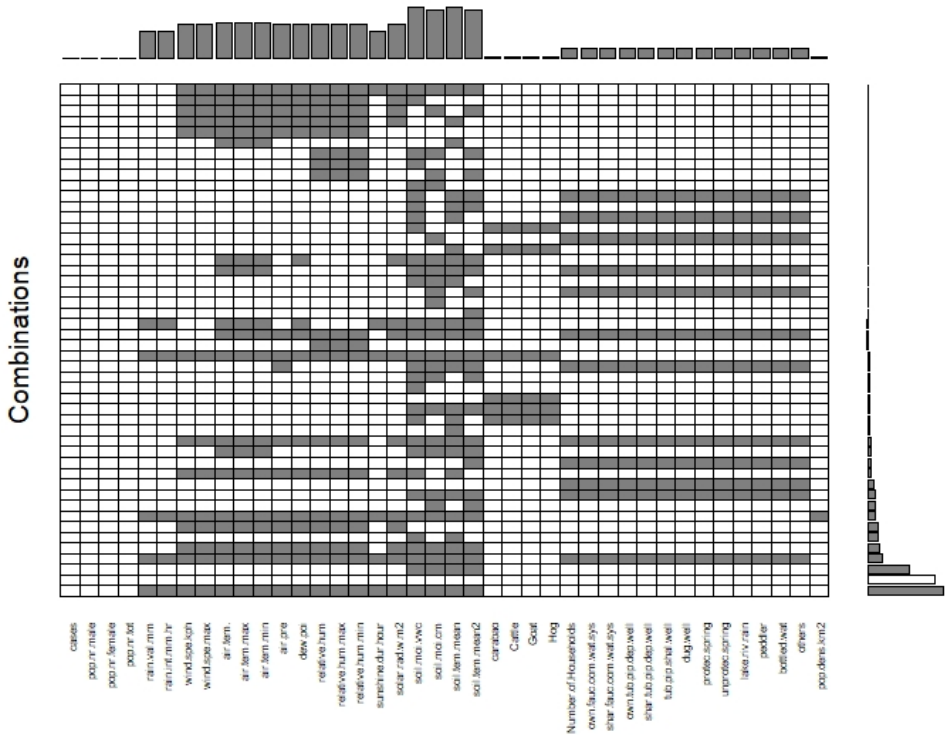



Figure 2: Aggregation plot of all variables.

Following the suspicions from the Appendix Table 4, one can observe very similar patterns of missingness within the weather and soil variables. Furthermore, animal farm and toilet & water availability traits have identical patterns in terms of non-available information. Keeping in mind that this study is done under the assumption that all variables are Missing Not at Random (MNAR), the

results observed earlier on in this section led to the following categorization of the data set features within INFORM (De Groeve et al., 2016):

- weather related variables can be categorized as *Independent Missing Variables* (since they are part of the natural phenomena and are out of human control) and can be found under the Infectious Hazards & Exposure dimension,
- the soil features can be a consequence of weather hence they will be considered as *Consequential Missing Variables* and are part of the Vulnerability dimension,
- the other features: animal farm, toilet & water availability and population density are under human influence hence they will be classified as *Regular Missing Variables* and should be linked to the dimensions of Lack of Coping Capacity and Infectious Hazards & Exposure (human subcategory).

For the remainder of this paper they will be referred to as IMV, CMV and RMV respectively.

A further complication of the data is the presence of multicollinearity. In order to see the magnitude of the problem, the Variance Inflation Factor (VIF) is analyzed. This is a method where each predictor variable is regressed on the remaining predictors. The resulting $R_j^2$ of the regression of variable $x_j$ on all others $x_{-j}$ is used in the computation of the VIF of said variable ($\text{VIF}_j$) the following way (Marquardt, 1970):

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{1}$$

It has been proven that a problematic variable will lead to a VIF value larger than 10 (Curto and Pinto, 2011). Upon a check of the current data set, the VIF-test points towards possible multicollinearity due to the following regressors: *pop.nr.male, pop.nr.female, wind.spe.kph, air.tem., air.tem.max, air.tem.min,relative.hum, relative.hum.max* and *relative.hum.min*. Full results of the VIF-test can be found in the Appendix Table 5.

The model predictive power or reliability is not reduced by multicollinearity. Calculations of individual predictors, however, are heavily affected. For example, in multicollinear data sets multivariate regressions cannot give valid results regarding which predictors are redundant as opposed to the others (Gujarati and Porter, 2003). Another phenomenon that occurs in multicollinear data sets is the fact that a small change in input information leads to a high change in model estimates (Belsley,

1991). Possible solutions include implementing Ridge Regressions because imposing size constraints alleviates the problem caused by multicollinearity. Positive large coefficients on a variable will be cancelled out through a negative large coefficient in its correlated counterpart (Hastie et al., 2005).

Considering that one of the most common tasks in econometrics is the choice of a parametric model which is able to fit some observations, it is necessary to assess the fit of any chosen model. Usually, one can choose the parameters of the model such that the sample mean is approximately the same as the theoretical population mean. In some cases, if one considers higher moments, it can be seen that observed variances fail to match the theoretical ones. Thus, when the observed variance is higher than the one of the theoretical framework, a circumstance known as over-dispersion has occurred. Authors such as Burnham and Anderson (2002) and Hilbe (2007) define over-dispersion as a phenomenon which occurs when the variation in the model outcome is greater than that which is integrated in the distribution. Simply put, the variance of the distribution used for modelling is in fact higher than expected. Over-dispersion may arise due to missing covariates, partial dependence or parameter heterogeneity.

Testing for over-dispersion can be done by checking the theoretical distribution variance against the actual. For example, in Poisson models, under the null hypothesis of equi-dispersion $H_0 : var(\mu) = \mu$, the alternative states that $H_A : var(\mu) = \phi\mu$. The over-dispersion coefficient is estimated through an ordinary least squares regression with no intercept of the form: $var(\mu) = \phi\mu$. Since under the null, the coefficient is standard normally distributed, a t-test on $\phi$ determines its significance. Having performed this test on the full current data set, a $p - value < 2.2e - 16$ with $z = 10.716$ has proven that the null hypothesis of equi-dispersion is rejected. The estimated value for the over-dispersion coefficient is $\phi = 177.1467$. This proof of the presence of non-equi-dispersion complicates model selection techniques. It will be shown in further sections that regular model selection techniques such as those based on the AIC cannot be performed on overly-dispersed data sets. As a result, Elastic Net Model Selection will be performed.

# 3 Literature

Current literature dives both into epidemic risk modelling as well as into imputation methods in epidemiological studies. This section will provide information on previous work on how missing data has been handled before by other authors vs. how it is handled now, as well as an overview of previous epidemic risk modelling as opposed to what econometric knowledge is used to investigate the current data set.

## 3.1 Imputation Methods

Previous studies opt for using one missing data handling method throughout, may this be Complete Case Analysis (CCA) (Liu and De, 2015), simple full kNN Imputation (Hierink, 2018) or Multiple Imputation via Model Based Imputation techniques (Harel et al., 2017). There have been some attempts at a Fully Conditional Specification, but as Liu and De (2015) state, these are rarely used. However, data sets with a proportion of missing values per variable of around 50% such as the one investigated here pose an interesting dilemma. There are simply too many parameters with missing values for a full Model Based Imputation (Kim et al., 2015), while the full kNN seems to naively overlook the intricate relationships between variables (Beretta and Santaniello, 2016).

This study proposes that a combination of imputation techniques should lead to less bias in the results. The question then becomes: which variables should be treated in which way. The framework of the INFORM index (De Groeve et al., 2016) is used as a guideline for this decision. INFORM defines what the European Commission selected as factors which contribute to risk for humanitarian crises and disasters. Figure 3 showcases the three dimensions with their respective categories and sub-components. The inherent nature of these dimensions (which contributed to them being separated the way the European Commission saw fit to do) is the basis for the separation of data sets into types of variables with their respective imputation treatment.

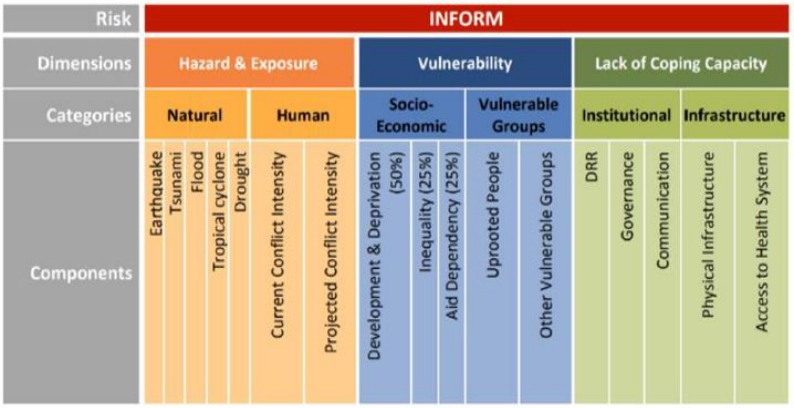



Figure 3: Risk for humanitarian crises and disasters index.

Starting with the first dimension, Hazard & Exposure variables, specifically the natural category, can be regarded as variables with independent effect. Furthermore, such variables are highly correlated with geographical aspects since "near things are more related than distant things" (Tobler, 1979). This leads to the best imputation method being k-Nearest-Neighbours (kNN) Imputation, a powerful donor-based imputation method (Altman, 1992). kNN's properties such as the lack of assumption on the distributions as well as its robustness provide the perfect environment for imputing naturally occurring types of variables with missingness. This type of imputation however is not suitable for multiple imputations exactly because it does not draw the imputed variables from a distribution. The present paper contributes to the usual kNN Imputation by separating the data set per geographical regions.

For the second dimension, Vulnerability can be seen as a consequence of other variables, thus variables with missing values within this dimension can be regarded as consequential therefore they can be modelled. Model Based Imputation techniques have been developed exactly for these purposes. Yohai et al. (1987) worked in creating estimations for linear regressions, Cantoni and Ronchetti (2001) developed robust techniques for estimating Logistic and Poisson regressors and Templ et al. (2011) tackled estimation in the presence of outliers. Through use of statistical models to estimate missing values and making a conditional model for each variable with missingess, the algorithms use repeated passes until convergence is achieved.

Variables which describe Lack of Coping Capacity seem to be directly influenced by human activity, much like the human category within the Hazard & Exposure dimension. As such, they

represent intricate causation patterns, hence a broader robust method is required. For this reason Fully Conditional Specification Imputation techniques have been developed. Buuren et al. (2006) used them for the first time as a general class of methods whose purpose is to specify imputations models as conditional distributions in multivariate data. They do not assume normality or linearity. They also handle continuity and categorical values, so the assumptions required for most imputation models can be relaxed for this case, leading to a broader spectrum of variables on which it can be applied (Liu and De, 2015). The current data set contains high levels of multicollinearity. Fully Conditional Specifications with Regression Trees are able to handle this limitation (Parker, 2010). This combination is a never before used imputation technique in epidemic studies.

Imputed data sets contain limitations when it comes to modelling and model selection. In order to reduce bias, multiple imputations are required. Sterne et al. (2009) discuss this mathematical fact in the field of medical research. Issues such as the optimal number of imputations required for asymptotic efficiency are discussed by Rubin (1987) and Buuren and Groothuis-Oudshoorn (2010). Furthermore, following the imputations, diagnostic methods must be used to asses the validity of the complete data set(s). Stuart et al. (2009) and Kolmogorov (1933) put forward both graphical and numerical diagnostics.

### 3.2 Modelling

Having a correctly fully imputed data set, the modelling work may begin. Keeping in mind the problems that arise when dealing with epidemiological studies: underreporting, multicollinearity and measurement errors, a model selection method which imposes penalties is required. Such models are part of the greater family of Generalized Linear Models. Publications such as Dobson and Barnett (1990) and Venables and Ripley (2002) are in support of these types of estimations of relationships between dependent and independent variables. Friedman et al. (2010) puts forward an algorithm for estimating Generalized Linear Models with the addition to the objective function of convex penalties. This can either mean one type of penalty such as Lasso or Ridge or a combination of both types which is identified in the literature as Elastic Net. These penalties' main objectives are either variable selection or coefficient shrinkage.

The need for this type of selection is that when dealing with over-dispersed data sets, regular

model selection criteria such as AIC cannot be estimated (Kullback, 1997). As AIC compares the differences between the model and the fitted candidate, it depends on the properties of maximum likelihood estimators (Kim et al., 2013). In the case of over-dispersion, AIC fails to select the best fitting model, opting for the over-fitted case instead (Anderson et al., 1994). Thus, model selection based on Elastic Net regularization seems to be a good alternative, because the variable selection is done via the penalty.

Other types of currently available variable selection criteria include, but are not limited to, Quasi-AIC (Kim et al., 2013), Quasi-AICc (Lebreton et al., 1992), KIC and Quasi-KIC (Kullback, 1997). These however rely on quasi-likelihood estimations. Wedderburn (1974) describes these as likelihood functions which do not correspond to any probability distribution. In turn, a relationship between the variance and the mean is created where the variance itself is a function of mean. This means that consistent estimates depend on the mean condition being correctly specified (Baltagi, 2015). To work around this limitation, present research uses an Elastic Net as a model selection technique.

The two types of regularizations found within an Elastic Net are the Lasso (Tibshirani, 1996) and Ridge (Tikhonov, 1963). Lasso is a regression analysis methodology which, as its name states, imposes a "least absolute shrinkage and selection operator". Using as penalty the absolute term, it has the power to shrink coefficients all the way to zero, thus *deselecting* variables. The goal is an easy interpretable sparse solution. On the other hand, Ridge regression has a different aim. Since the penalty is made up of a square of the magnitude of coefficients, the Ridge method keeps in the model all regressors leading to variable shrinkage, but not selection. The true power of Elastic Nets comes from the combination of the two types of regularizations. The combined penalty is a convex function and the end results often outperforms just Lasso regularization when used on real-world data (Zou and Hastie, 2005). An Elastic Net combines strongly correlated predictors and then proceeds to either include the entire group or not in the model.

Possible downsides of the penalized regression are described by Kyung et al. (2010). The authors warn about the dominance of the Ridge penalty over the Lasso in data sets which contain high levels of multicollinearity. As the preliminary analysis into the current data set has shown, there is clear evidence of multicollinearity. Furthermore, Kyung et al. (2010) mention the lack of

consensus as to how correct standard errors should be computed in a penalized regression setting. The best solution found thus far is by computing standard errors via Generalized Lasso estimators (Kyung et al., 2010), but since the literature offers no conclusive unanimity on this topic, the current research will not display standard errors for the estimates of the Elastic Net Regression. What it aims to do instead, is use the selected variables from the Elastic Net algorithm and include them into a Quasi-Poisson Regression. This new model will contain correctly selected variables with reliable coefficients and standard errors. Quasi-Poisson Regression is something that has been attempted before. Authors such as Berk and MacDonald (2008) discuss how in a Poisson setting observed variance being higher than the mean is related to over-dispersion, while Ver Hoef and Boveng (2007) discuss which model best fits count data: Quasi-Poisson or Negative Binomial.

In order to attempt to improve predictive power, a separate Bayesian Neural Network model is implemented. Such a construct is capable of integrating uncertainty. It is also robust to issues such as being useful in small data sets and over-fitting (Wu et al., 2018). Using a two layer neural network as per Foresee and Hagan (1997) with initial weight assignment according to the algorithm by Nguyen and Widrow (1990) and a Gauss-Newton optimization algorithm, the Bayesian optimization of the regularization parameters is performed. For these purposes the R *brnn* package is used as per Pérez-Rodríguez et al. (2013).

# 4 Methodology

This section provides information on the methodology of the current research. Since the main objective is developing expert based weights for all indicators for different infectious diseases in data sets with much missing values, one must first consider a statistically sound approach to filling in the non-available information. This is followed by appropriate modelling techniques as well as an implementation of a known better performing prediction framework. All codes used for the purposes of this research can be found at https://github.com/mufinel/master_thesis.

## 4.1 Pattern Based Imputation

As previously mentioned, the data will be separated into three categories as per INFORM (De Groeve et al., 2016). This leads to a sub-objective: recommending types of variable imputation given the category that the given set of variables find themselves in. The technique of assigning each missingness pattern to a type of variable within the INFORM index and then using a fitting imputation per category of variables represents one of the contributions this research brings to current literature: a Pattern Based Imputation method.

### 4.1.1 Independent Missing Variables

Thus, this paper begins with IMV type variables: the weather features. Their independence from human activity in combination with their geographical factor makes them good candidates for donor based imputation methods. Since natural phenomena occur inherently and tend to be more similar the closer the regions are to each other (Tobler, 1979), it seems to make sense that imputation via k-Nearest-Neighbours (kNN) should be the first candidate. However, including a geographical component in the imputation step leads to lower bias. The Philippines is split into three main island groups: Luzon, Mindanao and Visayas. By forcing the kNN Imputation to occur solely within these regions, a better filling in of the missing information is obtained.

### 4.1.2 Consequential Missing Variables

The second type of variables is the CMV. These are represented by the soil variables. Soil moisture, humidity and temperature is dictated in big part by the weather. Thus, one would expect that

soil variables are a function of the weather component. This consequential relationship led to the assumption that Model Based Imputation would fit this type of interconnected relationship best. For each variable within the soil type of features, a statistical model will be constructed. The algorithm is as follows (Alfons, 2019):

**Algorithm 1** Model Based Imputation Algorithm

Initialize missing values via a single imputation method - kNN

For each variable $x_j$ with $j = 1, .., p$

- Estimation of $x_j$ as response variable
- **For** each observation $x_i$ where $i = 1, .., n$
  - **If** $x_{ij}$ is missing, set $x_{ij}$ as the prediction from estimation model

Repeat until convergence

Return imputed data

To reduce bias multiple imputations are required. For this purpose Algorithm 1 will be repeated 5 times in order to obtain an asymptotic efficiency of 91% (Rubin, 1987). The final output will be constructed as the pooled result of all imputations.

#### 4.1.3 Regular Missing Variables

The final set of variables contains unknown underlying relationships between the human component and what the missing values could be. The assumption of the data being MNAR (Missing Not at Random) leads to only two possible approaches: disregarding the observations with missing values altogether or finding a robust method to impute them. In the data set present in the study both approaches are used. Since missing observations within animal farm animals and population density variables represent a total of 5% of the data set, this number can be considered small enough for them to be excluded from the analysis. The toilet & water availability traits however pose an interesting difficulty. Their effect on the world is not independent of human interaction, they are not a consequence of a naturally occurring event and they hold 12.5% of the missingness of the total observations in the data set. This rules out excluding them altogether. It can be seen from Figure 4 that they are made up of continuous variables which do not display a normal distribution.

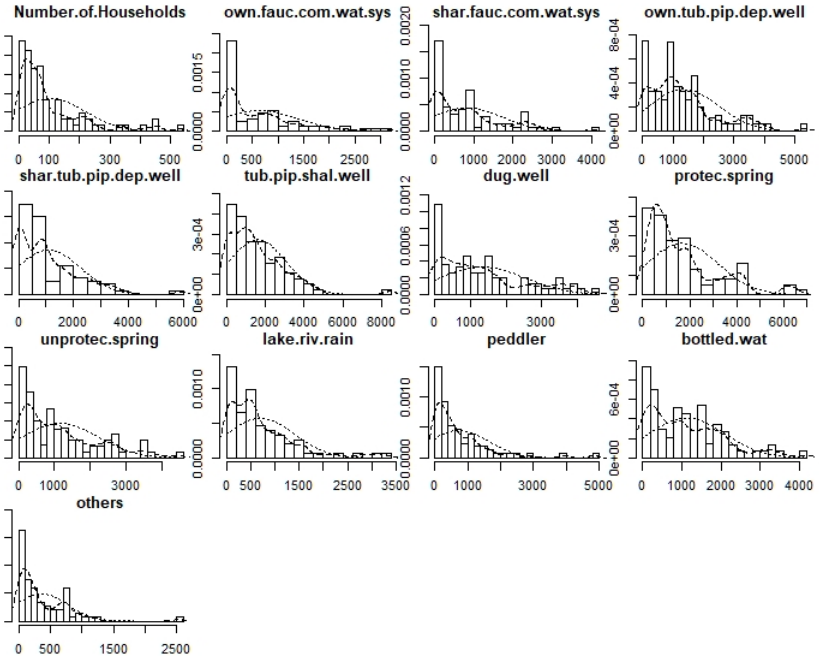



Figure 4: Histogram of RMV variables.

All of these lead to the conclusion that the most statistically sound method is multiple imputation by Fully Conditional Specification (FCS). This method is rarely ever used in epidemiologic studies. A reason for this is provided by Liu and De (2015) who state that little practical guidance as well as lack in availability of easy implementation and evaluation via this method could be the reasons why scientists in the epidemics field have held off from using it. However, due to its inclusion of continuous variables it is the best method for the problem at hand.

Multivariate FCS is an imputation technique which handles the data on a variable-by-variable basis. It uses a separate conditional density for each incomplete variable. Let $Z$ be the partially observed complete sample (thus including the dependent variable) of $p$ features with multivariate distribution $P(Z|\theta)$ completely specified by $\theta$, a vector consisting of unknown parameters divided into $p$ blocks. The posterior distribution of this vector is obtained by sampling iteratively from conditional distributions formed the following way:

$$\begin{aligned} &P(Z_1|Z_2, Z_3, ..., Z_p; \theta_1) \\ &\vdots \\ &P(Z_p|Z_1, Z_2, ..., Z_{p-1}; \theta_p) \end{aligned} \tag{2}$$

The parameters are not necessarily the product of some factorization of $P(Z|\theta)$, the true joint distribution. Furthermore they are treated as specific to their conditional densities. Having set starting values for $\theta^{(0)}$ and setting $t = 0$, the Gibbs sampler has the following form at the $t = T$-th iteration:

$$\begin{aligned}
\theta_1^{*(T)} &\sim P(\theta_1|z_1^{obs}, z_2^{(T-1)}, ..., z_p^{(T-1)}) \\
z_1^{*(T)} &\sim P(z_1^{miss}|z_1^{obs}, z_2^{(T-1)}, ..., z_p^{(T-1)}; \theta_1^{*(T)}) \\
&\vdots \\
\theta_p^{*(T)} &\sim P(\theta_p|z_p^{obs}, z_1^{(T-1)}, ..., z_{p-1}^{(T-1)}) \\
z_p^{*(T)} &\sim P(z_p^{miss}|z_p^{obs}, z_1^{(T-1)}, ..., z_{p-1}^{(T-1)}; \theta_p^{*(T)}),
\end{aligned} \tag{3}$$

where $z_j^{(T)} = (z_j^{obs}, z_j^{*(T)})$ is the result of the imputation of variable $j$ at iteration $T$. The cycle is repeated until convergence, after which the draws are considered as the first set of imputed values. For multiple imputation to be obtained, multiple such number of imputations must be achieved hence the cycle repeats itself a predefined $r$ number of times. Typically $r = [5, 10]$ 'complete' data sets are used (Buuren et al., 2006). The present research opts for 5 for ease of computation time.

As stated previously in the Data section, multicollinearity seems to plague the current data set. This is somewhat expected as the variables sometimes describe the same types of features. General Fully Conditional Specification techniques (such as predictive mean matching) do not work in such cases as the algorithm fails to converge (Huque et al., 2018). For this reason, the current paper uses Regression Trees instead of probability functions as it solves the interaction problem (Doove et al., 2014). For this purpose the R *mice* package with *cart* method was implemented (Buuren and Groothuis-Oudshoorn, 2010).

### 4.2 Diagnostics

The main assumption is that the data is MNAR (Missing Not At Random). Imputation methods can handle this type of data, but bias is regrettably introduced no matter what method is used (Gelman and Hill, 2006). To asses the accuracy of the imputations, diagnostics must be implemented. Stuart et al. (2009) tackle this issue for multiple imputation methods. They present both graphic and numeric diagnostics. The main obstacle when attempting this type of analysis is that the differences between observed and imputed observations does not necessarily imply a problem. In some cases, especially when the data is not missing at random, these differences could be exactly

what the imputation is trying to address. If, for example, the data set does not contain information in provinces with high poverty, the variables such as availability of toilet & water sources could be influenced by the factor of missingness leading to different distributions between the observed vs. the non-observed values. Field knowledge is the key to determining whether the imputations are reasonable or not.

Graphic diagnostics consist of a series of comparisons between the observed vs. non-observed distributions through histograms, quantile-quantile plots and density plots (Stuart et al., 2009) . Here, differences between distributions should be regarded with a critical eye. When assessing larger numbers of variables, however, graphical diagnostics may become difficult. Numeric diagnostics step in as a possible solution. Using specific measures, variables with large differences between the observed and imputed values will be selected. Such measures include, but are not limited to:

1. *z-test*: a difference in absolute terms of means of the imputed vs. observed values larger than two standard deviations away (Stuart et al., 2009),
2. *variance ratio test*: a ratio of variance smaller than 0.5 or larger than 2 in observed vs. non-observed values (Stuart et al., 2009),
3. *Kolmogorov-Smirnov test*: a non-parametric test done by comparing the equality of one-dimensional distributions of imputed and observed values (Kolmogorov, 1933), (Smirnov, 1948).

In general epidemic studies, authors use only one type of imputation for the entire data set, may it be multiple imputation via Model Based Imputation or simple kNN Imputation ((Harel et al., 2017) (Hierink, 2018)). The Red Cross have opted for a full kNN Imputation (from now on referred to as Naive kNN) on the entire data set in their studies (Hierink, 2018). For this reason a comparison between the current Pattern Based Imputation versus the Naive kNN Imputation is performed. In the Results section this paper will prove the superiority of the Pattern Based Imputation versus the Naive in all three types of variables. A simple comparison between the two methods is done via Kolmogorov-Smirnov testing as well as through graphical displays of the density plots (Kolmogorov, 1933). A comparison between the Pattern Based Imputation and Model Based Imputation has been attempted, but due to the high number of parameters and observations, the algorithm of the Model Based Imputation could not converge. Too many weights had to be estimated and furthermore rank-deficient fits seemed to overpower the predictions. A comparison of the sort may be of further

interest in future studies.

### 4.3 Modelling

In order to begin the modelling work, scaling must be imposed on the explanatory variables. This reduces the risk of bias in estimators. The current research opted for the following scaling technique: every variable with a mean higher than 100 is scaled down such that its mean is smaller or equal to 10.

Heavily imputed data sets introduce a delicate situation. Complications in both model building as well as model selection are part of the limitations of the multiple imputation technique. The bias introduced by imputing MNAR type variables is also of concern. This study proposes three methods for modelling new monthly cases of Dengue: one designed to select which regressors influence the dependent variable, one to explain how they affect presence of the disease and another for predictions. In order to compare forecasting power, the data set was split into a training and testing set via the $80\% - 20\%$ ruling (80% of total data set as training resulting in 786 observations while the remaining 126 represent the other 20% and is set as testing). Model fits were compared via $R^2$ values while predictions were compared to the actual values via the Mean Arctangent Absolute Percentage Error (MAAPE). Traditional methods such as the Mean Absolute Percentage Error (MAPE) cannot be applied in the present case because of its inability to handle actual values of zero (Hyndman and Koehler, 2006).

The MAAPE was selected as a measure of forecasting accuracy because of its ability to handle variables with different units, much like the traditional MAPE. What is more, this measure is able to handle cases where the dependent variable is zero, a complication of percentage error measures which are infinite or undefined in such situations. While regular MAPE calculates the slope as a ratio, this new technique introduced by Kim and Kim (2016) uses the slope as an angle:

$$MAAPE = \frac{1}{J}\sum_{j=1}^{J} \arctan\left(\left|\frac{A_j - F_j}{A_j}\right|\right). \tag{4}$$

This transformation switches the problem from the slope being a ratio of $|\frac{A-F}{A}|$, which ranges on $[0, \infty]$, to a slope as an angle $\theta = \arctan|\frac{A-F}{A}|$. Due to the properties of the arctangent function, $\theta$ can only vary on the interval $[0, \frac{\pi}{2}]$ , therefore no longer having the shortcoming of MAPE of

being unbounded for actual values close to zero. Furthermore, the MAAPE preserves the advantages of MAPE: scale-independence, easy interpretation and simple calculation. However, Kim and Kim (2016) do not recommend using this measure in cases where large errors may have business implications. This is due to the fact that MAAPE is much more robust to outliers than MAPE.

#### 4.3.1 Elastic Net Poisson Regression - Selection Model

The first approach includes the Selection Model. For this, a traditional statistical method combined with a regularization technique is used. Due to high over-dispersion, model selection based on the regular selection criteria such as AIC has proven to be impossible. This lead to model selection based on Lasso and Ridge penalty (also known as Elastic Net) being the better model selection method.

Elastic Net variable selection methods were introduced by Zou and Hastie (2005). The idea is that one can use together Lasso (Tibshirani, 1996) as well as Ridge (Hoerl and Kennard, 1988) regularization for model selection. In data sets with high inter-variable correlation, such as the present case, this method has high performance rates (Tibshirani, 1996).

Although both regularizations affect the coefficients, they do so in different ways. Lasso works by minimizing the residual sum of squares by penalizing the sum of the absolute value of all coefficients. This leads to the possibility that a coefficient $\beta_j$ could end up being equal to zero. The general Lasso model is:

$$\begin{aligned} \underset{\beta \in R^p}{\arg\min} \quad & \sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij}\right)^2 \\ \text{subject to} \quad & \sum_j |\beta_j| \leq v, \end{aligned} \tag{5}$$

where $\beta$ represent the coefficients, $\{y, x\}$ are the data and $v$ is a tuning parameter. Adding the penalty to an ordinary least squares (OLS) leads to the following loss function:

$$L_{Lasso}(\beta) = \sum_{i=1}^{N} (y_i - x_i'\beta)^2 + \lambda \sum_j |\beta_j|. \tag{6}$$

Ridge regression on the other hand works as coefficient shrinkage, not removal. It does so because the penalty is convex with loss function:

$$L_{Ridge}(\beta) = \sum_{i=1}^{N}(y_i - x_i'\beta)^2 + \lambda \sum_j \beta_j^2. \tag{7}$$

In both cases $\lambda$ is a tuning parameter. Furthermore, for all $j$, when $\lambda \uparrow \infty$, $\beta_j \downarrow$.

The power of using both regularization techniques comes from the simple reason that each method has its drawbacks which are then mitigated by the other. For example, in highly dimensional and inter-correlated data sets, Lasso begins to randomly select variables without concern of which variable fits best. In cases like this, Ridge outperforms Lasso (Tibshirani, 1996). This leads to the combined loss function:

$$\min_{\beta \in R^p} L_{elastic-net}(\hat{\beta}) = \min_{\beta \in R^p} \left[ \frac{1}{2N} \sum_{i=1}^{N}(y_i - x_i'\beta)^2 + \lambda \sum_j [(1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|)] \right], \tag{8}$$

where the $\alpha$ parameter decides the mix of Lasso vs. Ridge penalty.

The present research has as dependent variable a non-negative count as can be seen in Figure 5, hence a Poisson regression model seems appropriate.

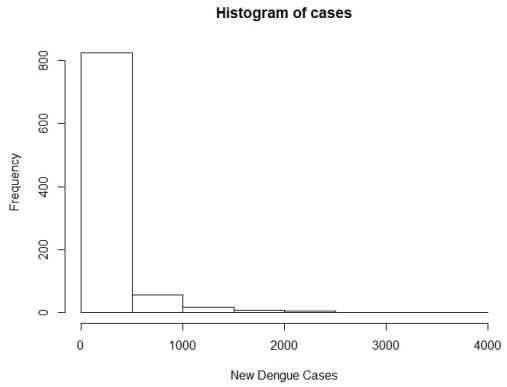



Figure 5: Histogram of the dependent variable: new Dengue cases.

As Friedman et al. (2010) state, the Poisson model is an exponential model where its positive mean is modelled on the log scale: $log_\mu(x) = \beta_0 + \beta_i' x$ hence $\{x_i, y_i\}$ have the log-likelihood function:

$$l(\beta|X,Y) = \sum_{i=1}^{N} \left( y_i(\beta_0 + \beta_i' x_i) - e^{\beta_0 + \beta_i' x_i} \right). \tag{9}$$

In order to obtain a model selection, a penalized log-likelihood is implemented. This uses coordinate descent in order to optimize the variable selection. The regularization path is computed at a grid of values for $\lambda$. The Elastic Net penalty $\alpha$ controls the gap between the Ridge ($\alpha = 0$, responsible for the shrinkage of correlated predictors towards each other) and Lasso ($\alpha = 1$, in charge of picking one predictor and discarding the others) while the $\lambda$ parameter is in control of the overall strength of the penalty. The $\alpha$ parameter is set to 0.5 such that equal weights are given to the two regularizations. In order to find the optimal $\lambda$ value, k-fold cross validation is implemented (with the default $k = 10$ folds). This method works by dividing the given data set into $k$ random groups of equal size. The procedure fits the model onto $k-1$ 'folds' and uses the remaining one as a validation set. Repeating this procedure for every separate group results in the mean cross-validated error (CVM):

$$CVM = \frac{1}{k}\sum_{i=1}^{k} MSE_i, \tag{10}$$

where $MSE_i$ is the Mean Squared Error for validation set $i$ (James et al., 2013). The *cv.glmnet* function within the *glmnet* package was used. This function returns two options for the optimal lambda: $\lambda_{1SE}$ and $\lambda_{min}$. The first represents the biggest $\lambda$ value such that the mean cross-validated error is one standard error away from the minimum CVM, while the second results in the minimum CVM error. The choice of which $\lambda$ should be used was done according to the principle of parsimony (Breiman et al., 1984). It will be shown that the 'one-standard-error' $\lambda$ model is not only parsimonious, but it also outperforms the 'minimum' $\lambda$ one in terms of model $R^2$. The minimization is thus as follows:

$$\min_{\beta} \; -\frac{1}{N} l(\beta|X,Y) + \lambda\Big((1-\alpha)\sum_{i=1}^{N}\frac{\beta_i^2}{2} + \alpha\sum_{i=1}^{N}|\beta_i|\Big). \tag{11}$$

#### 4.3.2 Quasi-Poisson Elastic Net Selected Model - Explanatory Model

This subsection presents the methodology used for the second modelling approach: the Explanatory Model. As stated before, in Elastic Net Regressions on highly correlated covariates the Ridge overpowers the Lasso leading to more coefficient shrinkage than necessary (Kyung et al., 2010). Furthermore, there is a lack of agreed upon reliable standard error estimation technique for Elastic Net coefficients (Kyung et al., 2010). In order to obtain consistent estimates both in terms of regressor weights as well as in terms of their significance, this research proposes a new technique: if one extracts the selected covariates from the Elastic Net Regression and feeds them to a model which can provide accurate estimations, consistent coefficients and variances are obtained.

The Poisson Regression is preferred as the dependent variable is a count of monthly new Dengue fever cases. Using the framework of Generalized Linear Models (GLM), the probabilistic Poisson model is as follows:

$$p(y|x,\beta) = \frac{e^{y\beta' x} e^{-e^{\beta' x}}}{y!}, \tag{12}$$

where $y$ is the dependent variable, $x$ are the dependent variables and $\beta$ represents the coefficients.

This framework however contains a rather important limitation. In Poisson modelling it is assumed that the variance is equal to the mean. When the data set contains over-dispersion, this assumption is limiting and leads to biased estimations of standard errors because, as mentioned before, the theoretical variance no longer matches the sample variance. One obtains correct variance estimations by using the Quasi-Poisson model which scales the variance by an over-dispersion parameter $\phi$:

$$var(\mu) = \phi\mu. \tag{13}$$

This leads to quasi-likelihood estimation methods put forward by Wedderburn (1974). The main effect on the estimation is when one calculates the standard errors. Under Poisson quasi-modelling, the regular Poisson standard errors are multiplied by $\sqrt{\hat{\phi}} = \sqrt{\frac{X^2}{N-P}}$ where $N$ is the number of individuals in the data set and $P$ is the total number of parameters. This correction accounts for over-dispersion.

Having already selected the optimal model regressors via the Elastic Net Poisson Regression, a Quasi-Poisson model on the selected covariates should provide reliable parameter calculations. The coefficients will no longer be highly shrunk towards zero (there no longer is any dominance of the Ridge penalty over the Lasso) and the variance estimation will be not only no longer limited (as Quasi-Poisson regressions does take into account over-dispersion), but also correctly estimated (as there is no need to correct for the Elastic Net penalty while evaluating variation).

This Quasi-Poisson Elastic Net Selected Model represents the contribution that paper offers to the current literature: a way to correctly estimate the standard errors as well as the coefficients in over-dispersed, highly imputed data sets after Elastic Net variable selection.

### 4.3.3 Bayesian Neural Network - Predictive Model

The third approach implies a combination of Bayesian Statistics and Neural Networks in order to obtain better predictive power. To understand what makes this approach such a powerful predictor, one must first grasp what Neural Networks are and how they operate. They can be thought of as a collection of interconnected nodes (often referred to as neurons) aggregated into layers where each connection sends a signal from one node to the other. Much like the synapses in the brain, information is thus propagated through the structure (McCulloch and Pitts, 1988). Figure 6 showcases the input layer, hidden layers and output layer. Each connection is given an initial weight while each neuron contains an activation function (usually an S-shaped function such as the sigmoid or tangent such that non-linearity is introduced in the model). The activation function determines a neuron's output according to the incoming information. Each node output is then propagated through the layers with further weights and activation function calculations (within the hidden layers) until the output layer is reached. The network learns via a process called backpropagation where the output is compared to the actual output from the training data. The weights are adjusted via retracing the steps back through the network and by taking into account an error minimization between the actual value and the neural network return.

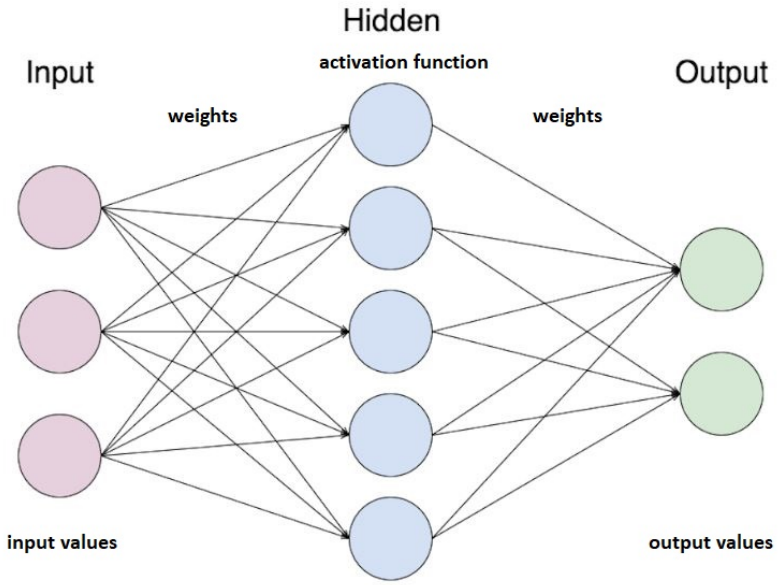



Figure 6: Neural Network

As stated above, conventional neural networks are trained via a process of minimizing the error

function which in itself may be derived via an underlying principle, for example the maximum likelihood. This approach contains limitations in the sense that finding the optimal model complexity which gives a good balance between dimension of the network and over-fitting is difficult. Bayesian Statistics aids the neural network optimization because it relaxes the algorithm. Furthermore, methods such as the error function minimization can be regarded as approximations of a Bayesian treatment (Bishop, 1997).

Dealing with uncertainty in a Bayesian setting is done via conditional probabilities: $P(x, y) = P(y|x)P(x)$ where $P(x)$ is the so called prior probability. This leads to what is known as Bayes' theorem:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}. \tag{14}$$

One can apply this to situations where a comparison between models is performed. Assuming models: $M_1, M_2, M_3$ with priors $P(M_1), P(M_2), P(M_3)$ (in case no known prior information is known then they are given equal prior probabilities) and data set $D$, the following holds for model $M_i$:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}. \tag{15}$$

The denominator $P(D)$ does not depend on $M_i$ hence the comparison between models can be done via $P(D|M_i)$ which MacKay (1991) have named the *evidence* for model $M_i$. In Regression problems one must estimate a parameter vector $w$ of weights and biases. Within the data set $D$ there are $N$ input vectors. Finding the distribution $p(w|D)$ is done similarly to how the conditional distribution of model $M_i$ has been achieved:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}. \tag{16}$$

The conventional way to find $w$ is by maximizing a likelihood function. In the Bayesian setting however this is done differently. An initial prior $p(w)$ is set such that it is a smooth and unconstraining function, usually the zero-mean Gaussian function:

$$p(w) = \frac{1}{Z_w(\alpha)} e^{-\frac{\alpha}{2}||w||^2}, \tag{17}$$

where $Z_w(\alpha) = (\frac{2\pi}{\alpha})^{\frac{w}{2}}$. The conditional data distribution $P(D|w)$ can be seen as a likelihood function in terms of $w$. The point where the posterior distribution reaches its maximum contains the vector $w_{mp}$ which is regarded as the most probable weight. Since the goal is the prediction of output values $y$ from new input values $x$, one needs a way to predict such values via a process of

integration over the weights:

$$p(y|x, D) = \int p(y|x, w)p(w|D)dw. \tag{18}$$

Ideally, one would use $w_{mp}$ in Equation 18. To determine such a vector one would need to seek the maximum of the posterior probability. It is easier however to find the minimum of the negative of the logarithm function (also referred to as the error function): $E(w)$. This depends on prior and likelihood and a weight decay term (Bishop, 1997). This error function can be approximated via a Taylor expansion at $w_{mp}$:

$$E(w) = E(w_{mp}) + \frac{1}{2}(w - w_{mp})^T H(w - w_{mp}), \tag{19}$$

where $H$ is the Hessian matrix.

The current research uses Foresee and Hagan (1997)'s algorithm for a Bayesian Neural Network optimization. Their research relies on MacKay (1991)'s framework for backpropagating networks. Given that a set of weight values $w$ is assigned to the node connections in the network, a mapping $y(x|w, M)$ is defined where $M$ is the architecture of the neural network (number of hidden layers, number of nodes per layer, activation functions etc.). The goal of any neural network algorithm is the minimization of some error function. In the present case this is the distance between the training set and the mapping. It is defined as follows on the data set $D$ of size $N$:

$$E_D(D|w, M) = \sum_{i=1}^{N}(y_i(x_i|w, M) - \hat{y}_i(x_i|w, M))^2, \tag{20}$$

where $y_i(x_i|w, M)$ is the model and $\hat{y}_i(x_i|w, M)$ is the neural network response. Foresee and Hagan (1997) define the model for $N$ individuals, $J$ variables and $K$ neurons as:

$$y_i(x_i|w, M) = g(x_i) + e_i = \sum_{k=1}^{K} w_k g_k\Big(b_k + \sum_{j=1}^{J} x_{ij}\beta_j^{(k)}\Big) + e_i, \tag{21}$$

where at neuron $k$: $w_k$ is the weight, $b_k$ the bias, $g_k(x) = (e^{2x-1})/(e^{2x+1})$ the activation function, $\beta_j^{(k)}$ the estimated parameter of the $j-$th variable in the network and $e_i$ the error term.

For better performance, an extra regularization term $E_W(w)$ is added to the objective function. This penalizes large weights such that smoother mappings can be created (MacKay, 1991). Is has been named *weight energy* and defined for the total number of weights and biases $B$ as:

$$E_W(w|M) = \sum_{i=1}^{B} \frac{1}{2} w_i^2. \tag{22}$$

Thus an objective function with parameters $\alpha$ and $\beta$ is constructed the following way (Foresee and Hagan, 1997):

$$F = \beta * E_D(D|w, M) + \alpha * E_W(w|M). \tag{23}$$

Foresee and Hagan (1997) state that the parameters $\alpha$ and $\beta$ should be initialized to 0 while the weights should be initially set according to the weight initialization algorithm by Nguyen and Widrow (1990).

As shown previously in this section, a Bayesian analysis requires a likelihood, a prior and a regularization parameter which guarantees that the total probability sums up to 1. In the present case these are denoted as $P(D|w, \beta, M)$, $P(w|\alpha, M)$ and $P(D|\alpha, \beta, M)$ respectively. This way, the conditional density function for weights given the data set is defined as:

$$P(w|D, \alpha, \beta, M) = \frac{P(D|w, \beta, M)P(w|\alpha, M)}{P(D|\alpha, \beta, M)}. \tag{24}$$

Assuming the noise is normally distributed and the prior distribution is also of a normal type, the probability densities are as follows:

$$P(D|w, \beta, M) = \frac{1}{Z_D(\beta)} e^{-\beta E_D(D|w,M)}, \tag{25}$$

where $Z_D(\beta) = \left(\frac{\pi}{\beta}\right)^{N/2}$ and

$$P(w|\alpha, M) = \frac{1}{Z_W(\alpha)} e^{-\alpha E_W(w|M)}, \tag{26}$$

where $Z_W(\alpha) = \left(\frac{\pi}{\alpha}\right)^{B/2}$.

By substituting Equations 25 and 26 in Equation 24 one obtains:

$$P(w|D, \alpha, \beta, M) = \frac{\frac{1}{Z_D(\beta)} \frac{1}{Z_W(\alpha)} e^{-(\beta E_D(D|w,M) + \alpha E_W(w|M))}}{P(D|\alpha, \beta, M)} \propto \frac{1}{Z_F(\alpha, \beta)} e^{-F}, \tag{27}$$

where $Z_F(\alpha, \beta) = Z_D(\beta) Z_W(\alpha)$.

If one wishes to maximize the posterior probability for the weights in Equation 27 in order to obtain the before mentioned $w_{mp}$, one must minimize $F$, the objective function. This brings the algorithm to the parameter optimization stage.

The density function of parameters $\alpha$ and $\beta$ given the data and the architecture is according to Bayes's rule as follows:

$$P(\alpha, \beta|D, M) = \frac{P(D|\alpha, \beta, M)P(\alpha, \beta|M)}{P(D|M)}. \tag{28}$$

Assuming uniformly distributed prior $P(\alpha, \beta|M)$, maximization of Equation 28 is done by maximizing the likelihood $P(D|\alpha, \beta, M)$. It is easy to see that this is the so called regularization parameter in Equation 24. Solving for it leads to the following:

$$P(D|\alpha, \beta, M) = \frac{P(D|w, \beta, M)P(w|\alpha, M)}{P(w|D, \alpha, \beta, M)}, \tag{29}$$

which is equal to:

$$P(D|\alpha, \beta, M) = \frac{Z_F(\alpha, \beta)}{Z_D(\beta)Z_W(\alpha)}, \tag{30}$$

where the only unknown is $Z_F(\alpha, \beta)$. As mentioned before in this section, this can be estimated via a Taylor Series expansion of $F$ around the minimum point $w_{mp}$. Solving for $Z_F$ yields:

$$Z_F = (2\pi)^{B/2}|H_{mp}^{-1}|^{1/2}e^{-F(w_{mp})}, \tag{31}$$

where $H = \nabla^2 F(w_{mp}) \approx 2\beta J_{mp}^T J_{mp} + 2\alpha I_B$ is the objective function approximated hessian matrix (where $J_{mp}$ represents the training set error Jacobian matrix at minimum point $w_{mp}$) (Demuth et al., 2014). Solving for the values of the parameters by deriving with respect to $\alpha$ and $\beta$ in the logarithm of Equation 30 leads to the following estimates:

$$\alpha = \frac{\gamma}{2 * E_W(w_{mp})} \quad \text{and} \quad \beta = \frac{n - \gamma}{2 * E_D(w_{mp})}, \tag{32}$$

where $\gamma = B - 2\alpha_{mp}\,\mathrm{tr}(H_{mp})^{-1}$ represents the number of parameters to be used.(Foresee and Hagan, 1997)

By continuous iterations between taking steps in the direction of the objective function minimization $F(w_{mp})$ in Equation 23, calculation of the effective number of parameters $\gamma$ and re-estimation of $\alpha$ and $\beta$ according to Equation 32 until convergence one obtains a Bayesian optimization of the regularization parameters. This type of optimization is referred to in the literature as the Levenberg-Marquardt algorithm (Foresee and Hagan, 1997).

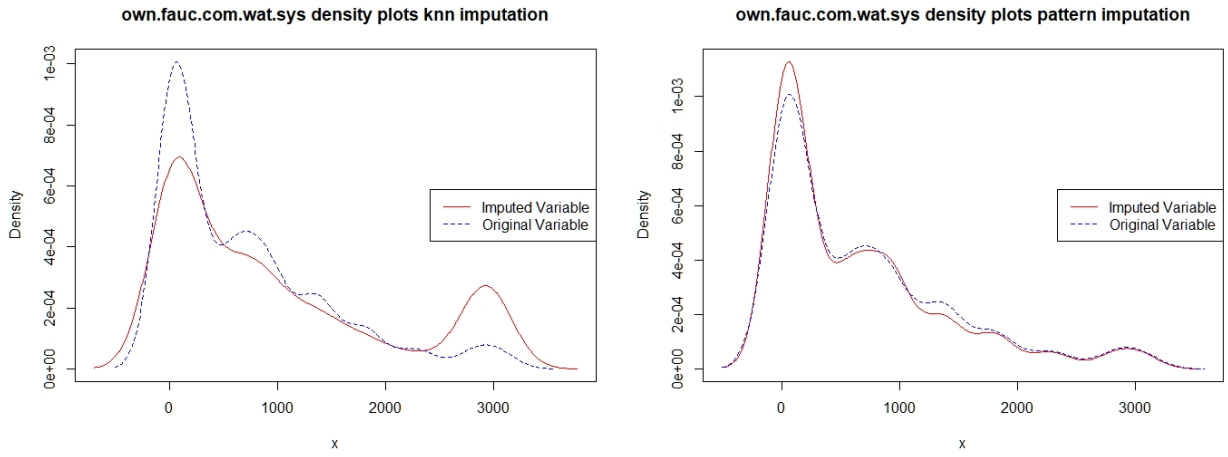
# 5 Results

This section contains the main results of the research. It is structured as follows: the first subsection compares the Pattern Based Imputation with the Naive kNN Imputation, the second and third subsections display the results of the Selection and Explanatory Models and the final subsection contains the results of the Predictive Model.

## 5.1 Imputation Comparison

In order to compare the imputation results, one variable from each of the three types is selected. From IMV, the comparison is done for *relative.hum.max* (maximum relative humidity), for CMV this study uses the *soil.moi.vwc* (soil moisture % VWC) and lastly for the RMV the *own.fauc.com.wat.sys* (own faucet and water system) variables are compared.

Figures 7, 8 and 9 display the results of Naive kNN Imputation and Pattern Based Imputation vs. the original density. The idea of this graphical diagnosis is to compare the original density of the variable (the blue dotted line) with the density after imputation (the red full line). If the two densities are very alike each other, than the imputation performed well. On the other hand, if the two shapes differ, then the imputation may be biased.

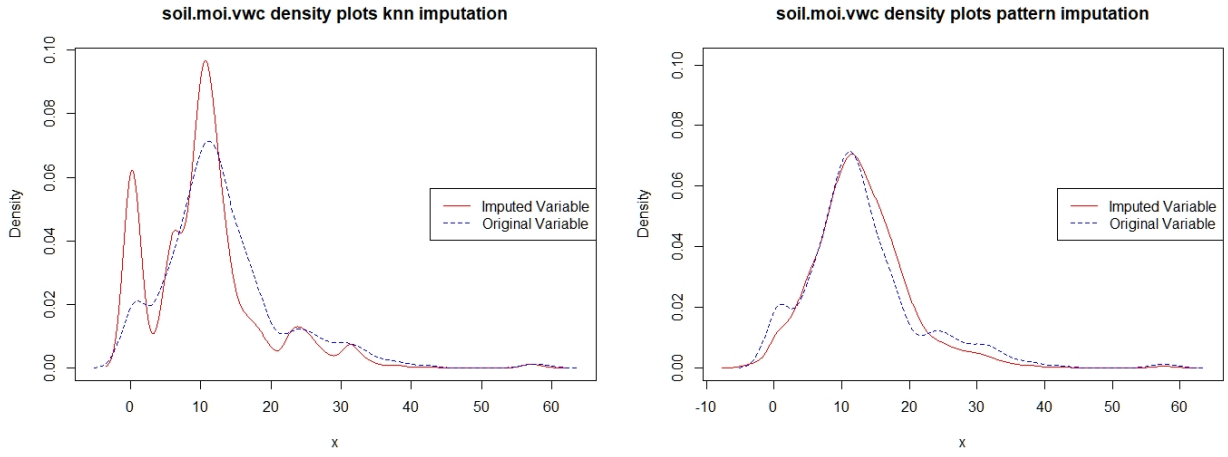



(a) Naive kNN Imputed vs Original.

(b) Pattern Imputed vs Original.

Figure 7: Density plots own faucet and water system before vs. after imputation.

It is clear the Pattern Based Imputation is superior to the Naive kNN. For example, in RMV type variables the two densities in Subfigure 7b are much closer than those of Subfigure 7a. Thus the density offered by the Pattern Based Imputation technique contains less bias.

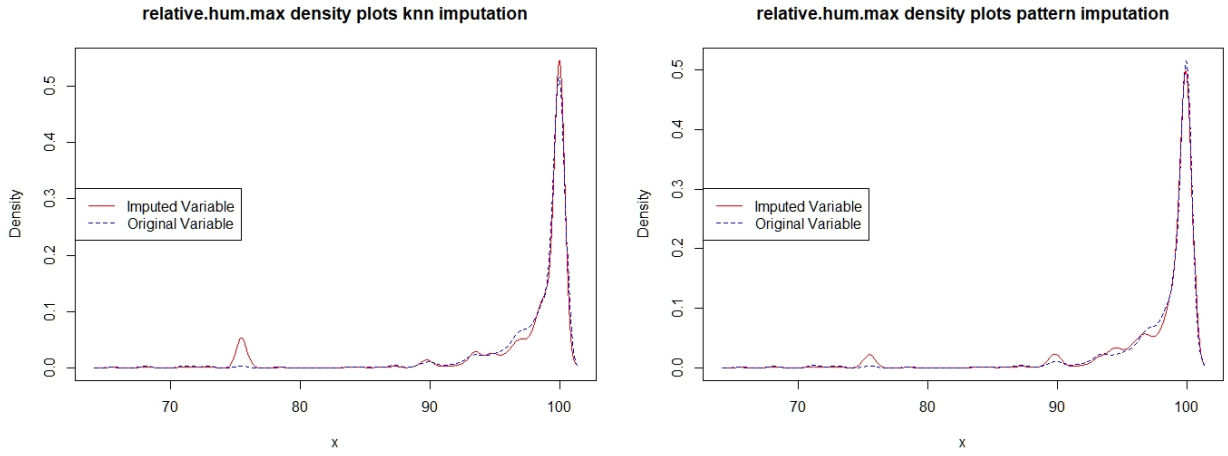



(a) Naive kNN Imputed vs Original.

(b) Pattern Imputed vs Original.

Figure 8: Density plots soil moisture before vs. after imputation.

The story repeats itself in the case of CMV variables as well. It can be seen that Subfigure 8b paints a closer resemblance between the Pattern Imputed data and the original as opposed to the situation in Subfigure 8a. It appears that Naive kNN Imputation favours $x = 0$ and $x = 10$ for imputation values. The Pattern Imputation follows the original distribution much closely, with no clear spikes for specific values of $x$.




(a) Naive kNN Imputed vs Original.

(b) Pattern Imputed vs Original.

Figure 9: Density plots relative humidity maximum before vs. after imputation.

A somewhat less striking difference can be found in the IMV variables. Figure 9 displays the differences in density between the Naive kNN and Pattern Imputation. The Pattern however still slightly matches the original more than that of the Naive. The reason why these two seem to be more similar however is that these weather variables were imputed using geographical kNN in the Pattern Based Imputation. The two methods are similar hence the results are alike. Geographical kNN however still slightly outperforms its Naive counterpart as Subfigure 9b does not contain the same high jump at $x$ value of around 75.

Furthermore, it can be seen from the Appendix Table 6 that the Naive kNN offers more variables whose Kolmogorov-Smirnov test is significant. This means that the sets of densities (the original and the post imputation) are sufficiently distinct for more variables than when one opts for the Pattern Based Imputation.

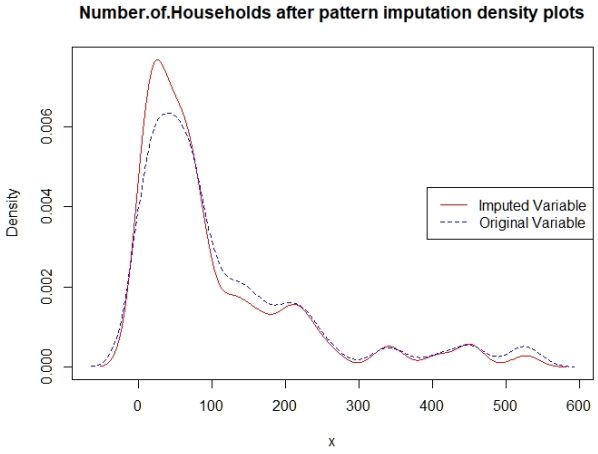
In order to diagnose the accuracy of the Pattern Based Imputation itself, three numeric methods were applied such that unusual variables can be flagged and further researched. In the Appendix Table 6 one can find the variables with significant Kolmogorov-Smirnov test. This means, according to the test, the original density and the density after imputation are significantly not equal for the following covariates: *rain.val.mm* (0.012), *wind.spe.kph* (0.005), *relative.hum.min* (0.016), *Number.of.Households* (0.025), *wind.spe.max* (0.000), *soil.tem.mean* (0.001) and *soil.moi.cm* (0.001). [1]

Furthermore, to these variables a z-test on the absolute difference in means was applied. If the z-value is greater than the critical value at significance of 0.05 ($z_{critical} = 1.96$), then the null hypothesis is rejected, thus the two means are significantly different. Appendix Table 6 shows that according to this test, none of the variables display odd behaviour, the means of the imputed and the non-imputed being not significantly different.

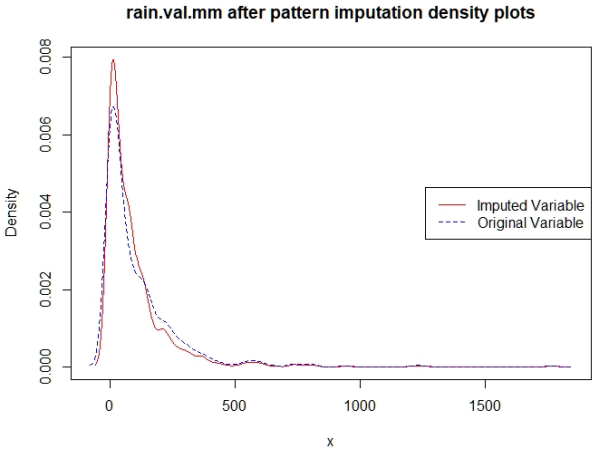
A last test is that of the variance ratio. If the first ratio (original/imputed) or the second ratio (imputed/original) in the Appendix Table 6 are less than 0.5 or greater than 2, then the imputed and the original variable are considered different enough. This test flagged the *wind.spe.kph* (wind speed in kmph) variable as the variance ratios were: 0.49 and 2.01 respectively.

[1] $p - value$ in brackets

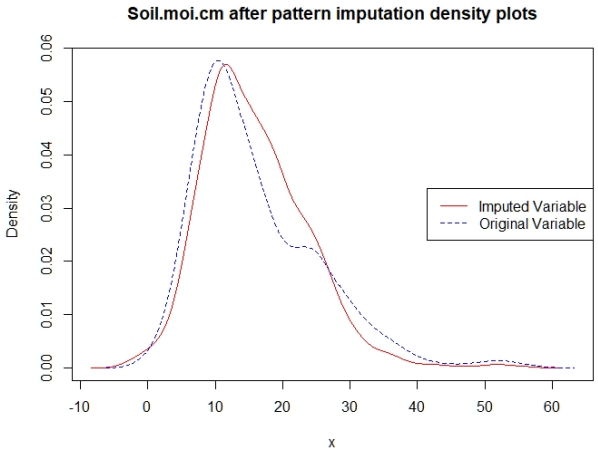
Having flagged the unusual variables with numeric tests, a graphical test on the densities of the imputed flagged variables may begin. Figure 10 showcases these densities as opposed to the original densities of the variables.



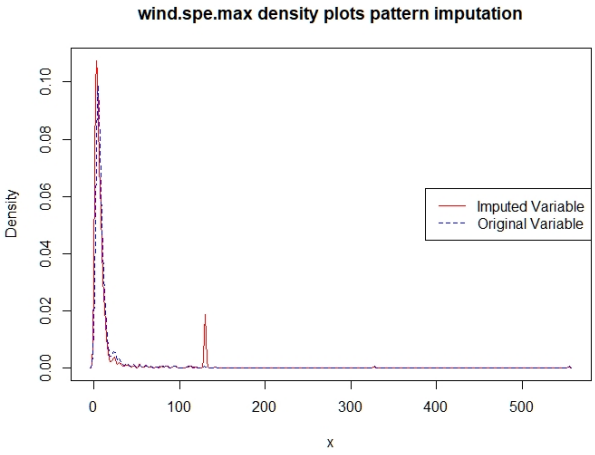

(a) Number of households.



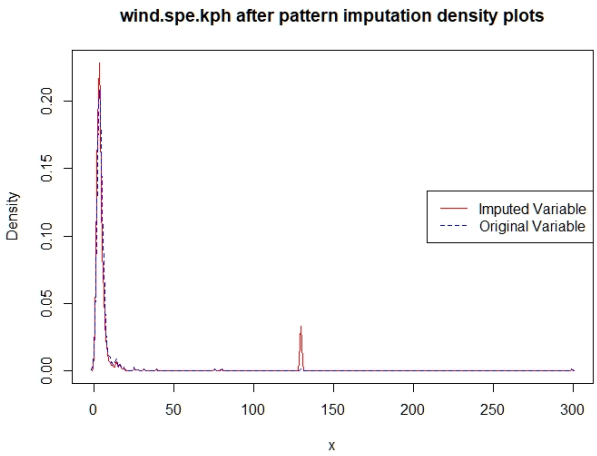

(b) Rain value per mm.



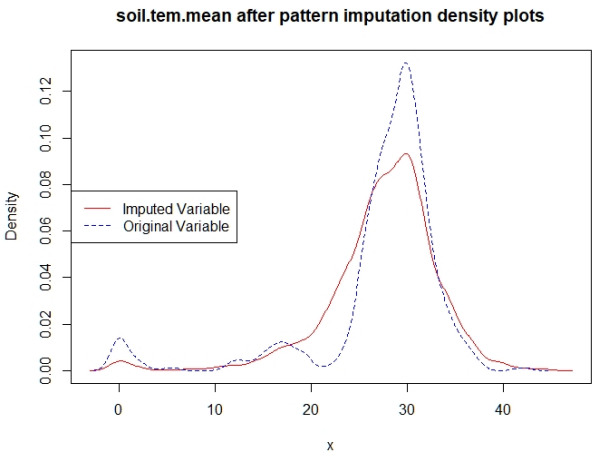

(c) Soil moisture.



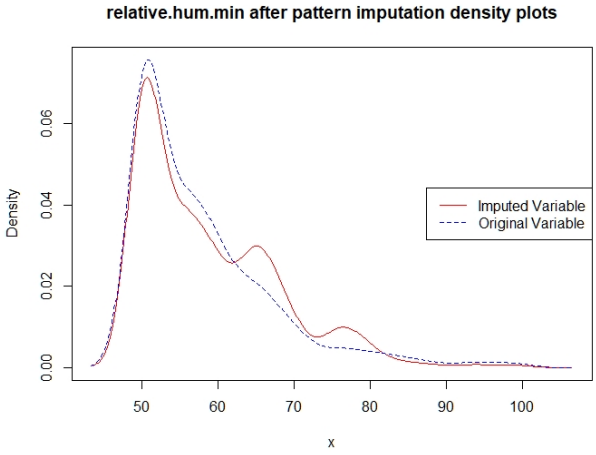

(d) Maximum wind speed.




(e) Wind speed in kmph.




(f) Mean soil temperature.




(g) Minimum relative humidity

Figure 10: Density plots flagged variables Pattern Imputed vs Original.

It is easy to see that even though the distributions are not identical, they match each other quite well. The imputed density and the original generally follow similar patterns as can be seen from the Figure 10. Subfigures 10d and 10f represent the variables with the most striking differences in outlines. The variable *wind.spe.kph* poses a jump at $x \approx 125$ and this could be the reason why this particular variable has been flagged by more numerical tests than the other regressors. The imputation of the variable *soil.tem.mean* from Subfigure 10f appears to underestimate the amount of times that a temperature of around 30 degrees is possible. However, the current study finds no reason to be alarmed by these small changes in the shapes of densities because the overall imputed and original outlines seem to be a close match for one another.

### 5.2 Elastic Net Poisson Regression - Selection Model

As previously mentioned, the data is plagued by problems such as multicollinearity and bias. The final impediment in model building is over-dispersion. The test shows a factor $\phi = 177.1467$ with p-value $= 2.2e-16$ hence there is evidence of over-dispersion. This means that the Poisson model selection cannot be performed via AIC selection. For this purpose the Elastic Net is used for variable selection.

The variables selected into the model are the following: *pop.nr.male, pop.nr.female, rain.val.mm, air.tem.max, cattle, own.fauc.com.wat.sys, own.tub.pip.dep.well, unprotec.spring, lake.riv.rain, bottled.wat,* and *pop.dens.km2*. The Appendix Table 7 displays this selection as well as the coefficients for each variable. These results are done with a $\lambda = \lambda_{1SE} = 50.08325$ since this led to the highest $R^2_{\lambda_{1SE}} = 0.1240878$ as opposed to the one via the minimum value for $\lambda = \lambda_{min} = 11.30389$ whose $R^2_{\lambda_{min}} = 0.1055164$.

### 5.3 Quasi-Poisson Elastic Net Selected Model - Explanatory Model

Having selected the variables which best fit the data distribution, this study uses a Quasi-Poisson Regression Model on the Elastic Net Selected variables (QPENselect) in order to obtain consistent coefficients and standard errors.

Table 1 showcases the results of the regression. The *Transformed Estimate* column is added for ease of interpretation. It changes the explanation of a variable effect from the expected log count of $y$ for a one-unit increase in $x$ to percentage of change in the $y$ for every unit increase of $x$. This is achieved by taking the exponent of all estimates (Long and Freese, 2006).

It appears that variables such as *pop.nr.male, rain.val.mm, air.temp.max, cattle* and *pop.dens.km2* positively influence the new cases of Dengue per month. On the other hand, *pop.nr.female, unprotec.spring, own.tub.pip.dep.well, own.fauc.com.wat.sys, bottled.wat* and *lake.riv.rain* negatively affect the monthly new cases of the disease. These results fall in line with what is known about Dengue fever. The disease bringing mosquito goes by the name of *Aedes aegypti* and require still waters in high temperatures in order to lay eggs (Ponnusamy et al., 2008). Hence an increase of one unit in maximum temperature and rain value per millimeter leads to an uptake in Dengue of

5.87% and 0.11% respectively. Furthermore, the higher the population density, the more risk of the disease being contracted. Hence an increase in one unit of *pop.dens.km2* leads to an increase of 0.06% in new cases of Dengue [2].

Differences in Dengue prevalence could be caused by gender-related distinct exposures such as time spent outside (Anker and Arima, 2011). Some previous studies have proven the tendency of greater Dengue incidence rate amongst men (Ooi, 2001). In the current study, this is suggested by the high coefficient carried by this variable. One extra unit of male population leads to an increase in Dengue risk of 273.68%. Furthermore, an additional unit of female population reduces the risk by 70.82%.

From the results it is also apparent that *cattle* has a positive influence on Dengue incidents. An extra unit of this animal type leads to an increase of 2.57% in the risk of contraction of the disease. Hasyim et al. (2018) have done research into how animal livestock influences the risk of malaria. They concluded that keeping such animals inside the house contributes to higher risk. This may be the case for the results found in the current study as well.

| | Estimate | Transformed Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|---|
| (Intercept) | 2.5074*** | 12.2731*** | 0.6205 | 4.0409 | 0.0001 |
| pop.nr.male | 1.3182*** | 3.7368*** | 0.3685 | 3.5774 | 0.0004 |
| pop.nr.female | -1.2317*** | 0.2918*** | 0.3752 | -3.2828 | 0.0011 |
| rain.val.mm | 0.0011*** | 1.0011*** | 0.0002 | 6.8927 | 0.0000 |
| air.tem.max | 0.0570 | 1.0587 | 0.0183 | 3.1095 | 0.0019 |
| cattle | 0.0254* | 1.0257* | 0.0152 | 1.6732 | 0.0947 |
| own.fauc.com.wat.sys | -0.0001 | 0.9999 | 0.0001 | -1.2969 | 0.1950 |
| own.tub.pip.dep.well | -0.0921 | 0.9120 | 0.0644 | -1.4309 | 0.1528 |
| unprotec.spring | -0.0876 | 0.9161 | 0.0593 | -1.4775 | 0.1399 |
| lake.riv.rain | -0.0001 | 0.9999 | 0.0001 | -1.3641 | 0.1729 |
| bottled.wat | -0.0232 | 0.9770 | 0.0583 | -0.3986 | 0.6903 |
| pop.dens.km2 | 0.0006*** | 1.0006*** | 0.0001 | 3.9697 | 0.0001 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | | | |

Table 1: Quasi-Poisson after Elastic Net Selected Model Results.

[2] All results are displayed *ceteris paribus*.

On the other hand, using sheltered water sources and water supply affects the breeding process of the disease carriers (Ponnusamy et al., 2008), hence owning such toilet and water facilities of one's own negatively affects the spread of the disease. One extra unit of an own faucet water system leads to a decrease in risk of 0.01% while an additional own tub and well reduces it by 8.80%. Having an extra bottled water measure also leads to a decline of 2.30% in risk.

Curiously, having one added unprotected spring unit leads to 8.39% less chance of contracting the disease. The reason why this happens is linked to the effect of the *lake.riv.rain* variable. For the latter, an extra unit of either lake, river or rain reduces the risk of Dengue by 0.01%. McNaughton et al. (2018) have studied the effect of water types on *Aedes aegypti*'s breeding grounds. The authors found that the mosquito does not breed in either of the following: creeks, lagoons, puddles, rivers or swamps. They also tend to only lay eggs in areas populated by humans. Springs, may they be protected or unprotected, as well as lakes and rivers represent remote areas where little to no humans have settled down.

The $R^2$ of the Quasi-Poisson with Elastic Net Selection is $R^2_{QPENselect} = 0.4738175$. This is much higher than just the Elastic Net Poisson Regression model whose $R^2$ was $R^2_{\lambda_{1SE}} = 0.1240878$. This proves the superiority of the new model in terms of fit.

### 5.4 Bayesian Neural Network - Predictive Model

The Bayesian Neural Network optimization provided the following parameters: $\gamma = 66.0618$, $\alpha = 0.2352$ and $\beta = 41.93$. It was constructed on two neurons with optimal scaling factor of 0.7006176. Table 2 showcases the coefficients of the variables selected by the Elastic Net in previous sections. The Appendix Table 8 contains the full results of the Bayesian Neural Network Analysis coefficients.

It can be seen that when looking at the coefficients of both neurons, most variables maintain their influence on the dependent variable: *pop.nr.male, rain.val.mm, air.tem.max, pop.dens.km2* lead to an increase in Dengue cases whereas an uptake in *own.fauc.com.wat.sys, lake.riv.rain, bottled.wat* decrease the risk. The difference in interpretation lies within the *pop.nr.female, cattle* and *own.tub.pip.dep.well* variables. Their coefficient from the first neuron still matches the results from

the previous sections, but the second neuron's effect leads to the opposite conclusion: *cattle* decreases the risk of Dengue and an extra unit of *pop.nr.female* and *own.tub.pop.dep.well* means an increase in chance of contracting the disease. Since the neurons have different conclusions regarding these variables, this study considers the differences in interpretations as opposed to the previous sections to be of no concern. Changing the architecture of the Neural Network may lead to more conclusive results in the case of the indecisive variables.

| Variables | Neuron 1 | Neuron 2 |
|---|---|---|
| weight | 3.090311 | 2.709838 |
| bias | 0.780622 | -0.23502 |
| pop.nr.male | 0.519813 | 0.083543 |
| pop.nr.female | -0.46333 | 0.052662 |
| rain.val.mm | 0.04555 | 3.915465 |
| air.tem.max | 0.037388 | 2.526867 |
| cattle | 0.043424 | -4.09695 |
| own.fauc.com.wat.sys | -0.00378 | -0.13511 |
| own.tub.pip.dep.well | -0.02127 | 1.221236 |
| unprotec.spring | 0.021065 | -1.6285 |
| lake.riv.rain | -0.01229 | -0.76821 |
| bottled.wat | 0.014114 | -0.73865 |
| pop.dens.km2 | 0.018748 | 1.130989 |

Table 2: Coefficients Bayesian Neural Network of Elastic Net Selected Variables.

Table 3 displays the MAAPE of all three models considered in this study. The predictions were done after a $80\% - 20\%$ split of the data which resulted in 126 observations in the testing set. Interpreting MAAPE results is easily done: the smaller the value of the mean arctangent absolute percentage error, the better the accuracy of the forecast. The Bayesian Neural Network outperforms the Selection Model in terms of forecasting, but fails to overcome the predictive power of the Quasi-Poisson Regression after Elastic Net Selected Model (QPENselect) whose mean arctangent percentage error rate is 70.48%.

| Model Name | MAAPE |
|---|---|
| Quasi-Poisson Regression<br>Elastic Net Selected Model | 0.7048382 |
| Bayesian Neural Network | 0.9017431 |
| Elastic Net Poisson Regression | 1.047515 |

Table 3: MAAPE comparison amongst all models.

This proves the superiority of the QPENselect in terms of forecasting power. The previous section proved its power in terms of fit. This study thus concludes the new modelling technique provides the best results given the common limitations found in epidemic studies: multicollinearity, bias from imputation and over-dispersion.

# 6 Conclusion

This study has not only tackled the problem of missing values, but it also gives a guideline of how such issues can be overcome by separating the variables into categories and tackling each group individually (IMV, CMV, RMV). It adds to the present literature not only this guideline, but also methods for geographical and conditional imputation. Furthermore, this study introduces a way of dealing with multicollinear unknown interconnected variables by using a Fully Conditional Specification Imputation method which relies on Regression Trees. It has been shown that this Pattern Based Imputation outperforms the Naive kNN Imputation (a method widely used in Epidemic studies, especially within institutions such as the Red Cross). Diagnostic testing has resulted in the validation of the Pattern Based Imputation. The imputed data sets matched the original ones in density as well as in variance and mean.

In order to maintain reliability, fitting modelling techniques for data sets with highly imputed values have been presented. The Selection Model (Elastic Net Poisson Regression) and the Predictive Model (Bayesian Neural Network) were compared to a new Explanatory technique: Quasi-Poisson Regression after Elastic Net Selection Model. This new approach outperformed both the selection as well as the Predictive Model according to model fit and forecasting power. This study suggests that when dealing with highly imputed data sets where there is unexplained residual variance as well as multicollinearity, Generalized Linear Models on penalized regression selected variables is the best approach. Due to the penalized model selection step, one is certain that this is the best fitting model based on the data set. Furthermore, it provides a stable coefficient and standard error estimation which simple Elastic Net Regression algorithms fail to do.

Further research into how Pattern Based Imputation Techniques perform as opposed to other imputation techniques should be investigated. Computation time as well as how the methods affect bias introduction represent possible research questions for the future. Additionally, better structured Bayesian Neural Networks with more intricate priors could lead to better predictions, thus outperforming the Quasi-Poisson Elastic Net Selected Model. Different variable selection techniques prior to the quasi-likelihood estimation could lead to higher performance. This study however has proven that given the difficult data set, the new modelling technique exceeded the simple Elastic Net Poisson Regression as well as the straightforward Bayesian Neural Network both in terms of model fit and forecasting power.

# References


Alfons, A. (2019). Missing data mechanisms and single imputation. Topics in Advanced Statistics lecture slides from Erasmus University Rotterdam.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Anderson, D., Burnham, K., and White, G. (1994). Aic model selection in overdispersed capture-recapture data. *Ecology*, 75(6):1780–1793.

Anker, M. and Arima, Y. (2011). Male–female differences in the number of reported incident dengue fever cases in six asian countries. *Western Pacific surveillance and response journal: WPSAR*, 2(2):17.

Baltagi, B. H. (2015). *The Oxford handbook of panel data*. Oxford Handbooks.

Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. Number 519.536 B452. Wiley New York.

Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, 16(3):74.

Berk, R. and MacDonald, J. M. (2008). Overdispersion and poisson regression. *Journal of Quantitative Criminology*, 24(3):269–284.

Bishop, C. M. (1997). Bayesian Neural Networks. *Journal of the Brazilian Computer Society*, 4.

Breiman, L., Friedman, J. H., Olsen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*.

Burnham, K. P. and Anderson, D. R. (2002). Model selection and.

Buuren, S. V., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030.

Curto, J. D. and Pinto, J. C. (2011). The corrected vif (cvif). *Journal of Applied Statistics*, 38(7):1499–1507.

De Groeve, T., Poljansek, K., and Vernaccini, L. (2016). Index for risk management—inform—concept and methodology. *Luxembourg: Publications Office of the European Union*.

Demuth, H. B., Beale, M. H., De Jess, O., and Hagan, M. T. (2014). *Neural Network Design*. Martin Hagan, USA, 2nd edition.

Dobson, A. J. and Barnett, A. G. (1990). *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Stat. Data Anal.*, 72:92–104.

Foresee, F. D. and Hagan, M. T. (1997). Gauss-newton approximation to bayesian learning. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 3, pages 1930–1935. IEEE.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gelman, A. and Hill, J. (2006). *Missing-data imputation*, page 529–544. Analytical Methods for Social Research. Cambridge University Press.

Gujarati, D. and Porter, D. (2003). Multicollinearity: What happens if the regressors are correlated. *Basic econometrics*, 363.

Harel, O., Mitchell, E. M., Perkins, N. J., Cole, S. R., Tchetgen Tchetgen, E. J., Sun, B., and Schisterman, E. F. (2017). Multiple Imputation for Incomplete Data in Epidemiologic Studies. *American Journal of Epidemiology*, 187(3):576–584.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Hasyim, H., Dhimal, M., Bauer, J., Montag, D., Groneberg, D. A., Kuch, U., and Müller, R. (2018).

Does livestock protect from malaria or facilitate malaria prevalence? a cross-sectional study in endemic rural areas of indonesia. *Malaria journal*, 17(1):302.

Hierink, F. (2018). The development of the epidemics risk and priority index: The environmental risk factors of epidemics.

Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge University Press.

Hoerl, A. and Kennard, R. (1988). Ridge regression. *Encyclopedia of Statistical Sciences*, 8:129–136.

Huque, M. H., Carlin, J. B., Simpson, J. A., and Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(1):168.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kekulé, A. (2015). Learning from ebola virus: How to prevent future epidemics. *Viruses*, 7.

Kim, H.-J., Cavanaugh, J., A. Dallas, T., and Foré, S. (2013). Model selection criteria for overdispersed data and their application to the characterization of a host-parasite relationship. *Environmental and Ecological Statistics*, 21.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Kim, S., Sugar, C. A., and Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in medicine*, 34(11):1876–1888.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412.

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological monographs*, 62(1):67–118.

Liu, Y. and De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International journal of statistics in medical research*, 4(3):287.

Long, J. S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

MacKay, D. J. (1991). Bayesian interpolation. *Neural Computation*, 4:415–447.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612.

McCulloch, W. S. and Pitts, W. H. (1988). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 52:99–115.

McNaughton, D., Miller, E., and Tsourtos, G. (2018). The importance of water typologies in lay entomologies of aedes aegypti habitat, breeding and dengue risk: A study from northern australia. *Tropical medicine and infectious disease*, 3(2):67.

Nguyen, D. and Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *1990 IJCNN International Joint Conference on Neural Networks*, pages 21–26. IEEE.

Ooi, E. E. (2001). Changing pattern of dengue transmission in singapore. *WHO Regional Office for South-East Asia*.

Parker, R. (2010). *Missing Data Problems in Machine Learning*. VDM Verlag.

Pérez-Rodríguez, P., Gianola, D., Weigel, K., Rosa, G., and Crossa, J. (2013). An r package for fitting bayesian regularized neural networks with applications in animal breeding. *Journal of Animal Science*, 91(8):3522–3531.

Ponnusamy, L., Xu, N., Nojima, S., Wesson, D. M., Schal, C., and Apperson, C. S. (2008). Identification of bacteria and bacteria-associated chemical cues that mediate oviposition site preferences by aedes aegypti. *Proceedings of the National Academy of Sciences*, 105(27):9262–9267.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):543–546.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393.

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009). Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, 169(9):1133–1139.

Templ, M., Alfons, A., and Filzmoser, P. (2012). Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1):29–47.

Templ, M., Kowarik, A., and Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793–2806.

The Netherlands Red Cross (2018). ERA: Epidemic Risk Assessment. *510 An Initiative Of The Netherlands Red Cross*, 2018, Accessed: 30-02-2019.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. In *Dokl. Akad. Nauk.*, volume 151, pages 1035–1038. Soviet Mathematics.

Tobler, W. R. (1979). Cellular geography. *Philosophy in Geography*, 20:379–386.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. (2018). Deterministic variational inference for robust bayesian neural networks.

Yohai, V. J. et al. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320.

# 7 Appendix

| Percentage NA value | Variable name | Variable description |
|---|---|---|
| 0 | cases | new monthly cases of Dengue fever |
| 0 | pop.nr.male | total male population |
| 0 | pop.nr.female | total female population |
| 0 | pop.nr.tot | total population |
| 0.340625 | rain.val.mm | rain value (mm) |
| 0.340625 | rain.int.mm.hr | rain intensity (mm/hr) |
| 0.435416 | wind.spe.kph | wind speed (kph) |
| 0.435416 | wind.spe.max | maximum wind speed (kph) |
| 0.455208 | air.tem. | air temperature (°C) |
| 0.455208 | air.tem.max | maximum air temperature (°C) |
| 0.455208 | air.tem.min | minimum air temperature (°C) |
| 0.442708 | air.pre | air pressure (hPa) |
| 0.44375 | dew.poi | dew point (°C) |
| 0.444791 | relative.hum | relative humidity (%) |
| 0.444791 | relative.hum.max | maximum relative humidity (%) |
| 0.444791 | relative.hum.min | minimum relative humidity (%) |
| 0.341666 | sunshine.dur.hour | sunshine duration hour(s) |
| 0.442708 | solar.rad.w.m2 | solar radiation (m2) |
| 0.654166 | soil.moi.vwc | soil moisture (% VWC) |
| 0.618756 | soil.moi.cm | soil moisture(30cm) (% VWC) |
| 0.65 | soil.tem.mean | soil temperature mean |
| 0.6260416 | soil.tem.mean2 | soil temperature mean squared |
| 0.025 | carabao | carabao |
| 0.025 | cattle | cattle |
| 0.025 | Goat | goat |
| 0.025 | Hog | hog |
| 0.125 | Number.of.Households | number of households |
| 0.125 | own.fauc.com.wat.sys | own use faucet community water system |
| 0.125 | shar.fauc.com.wat.sys | shared use faucet community water system |
| 0.125 | own.tub.pip.dep.well | own use tubed/piped deep well |
| 0.125 | shar.tub.pip.dep.well | shared use tubed/piped deep well |
| 0.125 | tub.pip.shal.well | tubed/piped shallow well |
| 0.125 | dug.well | dug well |
| 0.125 | protec.spring | protected spring |
| 0.125 | unprotec.spring | unprotected spring |
| 0.125 | lake.riv.rain | lake, river, rain |
| 0.125 | peddler | peddler |
| 0.125 | bottled.wat | bottled water |
| 0.125 | others | others |
| 0.025 | pop.dens.km2 | population density (per km2) |

Table 4: Percentage of missing values, variables and variable description.

| | VIF | Detection |
|---|---|---|
| pop.nr.male | 3742.199 | 1 |
| pop.nr.female | 3810.398 | 1 |
| rain.val.mm | 1.46592 | 0 |
| rain.int.mm.hr | 1.470992 | 0 |
| wind.spe.kph | 11.28316 | 1 |
| wind.spe.max | 6.174881 | 0 |
| air.tem. | 83.58987 | 1 |
| air.tem.max | 26.6226 | 1 |
| air.tem.min | 31.39782 | 1 |
| air.pre | 1.532903 | 0 |
| dew.poi | 1.465146 | 0 |
| relative.hum | 37.25954 | 1 |
| relative.hum.max | 13.47417 | 1 |
| relative.hum.min | 39.90851 | 1 |
| sunshine.dur.hour | 1.528944 | 0 |
| solar.rad.w.m2 | 2.960523 | 0 |
| soil.moi.vwc | 1.948302 | 0 |
| soil.moi.cm | 1.959306 | 0 |
| soil.tem.mean | 1.383714 | 0 |
| soil.tem.mean2 | 1.506878 | 0 |
| carabao | 2.902461 | 0 |
| cattle | 5.43727 | 0 |
| Goat | 6.15039 | 0 |
| Hog | 5.17202 | 0 |
| Number.of.Households | 9.056002 | 0 |
| own.fauc.com.wat.sys | 3.368912 | 0 |
| shar.fauc.com.wat.sys | 2.429458 | 0 |
| own.tub.pip.dep.well | 5.568337 | 0 |
| shar.tub.pip.dep.well | 3.107625 | 0 |
| tub.pip.shal.well | 5.512502 | 0 |
| dug.well | 5.717588 | 0 |
| protec.spring | 8.348004 | 0 |
| unprotec.spring | 4.793932 | 0 |
| lake.riv.rain | 2.61835 | 0 |
| peddler | 3.333385 | 0 |
| bottled.wat | 4.461705 | 0 |
| others | 4.555849 | 0 |
| pop.dens.km2 | 7.159102 | 0 |

Table 5: VIF results.

| Imputation type | Variable | p-value | z-test | Variance ratio orig/imp | Variance ratio imp/orig |
|---|---|---|---|---|---|
| Pattern based | rain.val.mm | 0.01271 | 0.960782 | 1.15945 | 0.862478 |
| | wind.spe.kph | 0.005578 | 0.847005 | 0.4967985 | 2.012888 |
| | relative.hum.min | 0.01661 | 0.211372 | 1.007598 | 0.992459 |
| | Number.of.Households | 0.02594 | 0.976566 | 1.100964 | 0.908295 |
| | wind.spe.max | 2.33E-06 | 0.423054 | 0.8668213 | 1.15364 |
| | soil.tem.mean | 0.001864 | 0.127458 | 1.177273 | 0.84942 |
| | soil.moi.cm | 0.01434 | 0.00364 | 1.174891 | 0.851143 |
| kNN based | rain.val.mm | 0.02294 | - | - | - |
| | rain.int.mm.hr | 0.008781 | - | - | - |
| | wind.spe.kph | 0.000838 | - | - | - |
| | wind.spe.max | 3.05E-14 | - | - | - |
| | air.pre | 0.03496 | - | - | - |
| | relative.hum.min | 0.03386 | - | - | - |
| | sunshine.dur.hour | 0.01631 | - | - | - |
| | soil.moi.vwc | 2.25E-07 | - | - | - |
| | Number.of.Households | 2.38E-05 | - | - | - |
| | own.fauc.com.wat.sys | 3.66E-06 | - | - | - |
| | shar.fauc.com.wat.sys | 2.47E-06 | - | - | - |
| | own.tub.pip.dep.well | 3.66E-06 | - | - | - |
| | shar.tub.pip.dep.well | 2.47E-06 | - | - | - |
| | tub.pip.shal.well | 2.47E-06 | - | - | - |
| | dug.well | 2.38E-05 | - | - | - |
| | protec.spring | 2.47E-06 | - | - | - |
| | unprotec.spring | 7.87E-06 | - | - | - |
| | lake.riv.rain | 2.47E-06 | - | - | - |
| | peddler | 2.47E-06 | - | - | - |
| | bottled.wat | 2.47E-06 | - | - | - |
| | others | 1.65E-05 | - | - | - |

Table 6: Results of tests on Pattern Based vs. Naive kNN Imputation.

| Variable Name | Coefficient | Transformed Coefficient |
|---|---|---|
| Intercept | 3.9268 | 50.7450 |
| pop.nr.male | 0.0450 | 1.0460 |
| pop.nr.female | 0.0446 | 1.0456 |
| rain.val.mm | 0.0007 | 1.0007 |
| rain.int.mm.hr | . | . |
| wind.spe.kph | . | . |
| wind.spe.max | . | . |
| air.tem. | . | . |
| air.tem.max | 0.0207 | 1.0209 |
| air.tem.min | . | . |
| air.pre | . | . |
| dew.poi | . | . |
| relative.hum | . | . |
| relative.hum.max | . | . |
| relative.hum.min | . | . |
| sunshine.dur.hour | . | . |
| solar.rad.w.m2 | . | . |
| soil.moi.vwc | . | . |
| soil.moi.cm | . | . |
| soil.tem.mean | . | . |
| soil.tem.mean2 | . | . |
| Carabao | . | . |
| cattle | 0.0118 | 1.0119 |
| Goat | . | . |
| Hog | . | . |
| Number.of.Households.y | . | . |
| own.fauc.com.wat.sys.y | 0.0000 | 1.0000 |
| shar.fauc.com.wat.sys.y | . | . |
| own.tub.pip.dep.well.y | -0.0240 | 0.9763 |
| shar.tub.pip.dep.well.y | . | . |
| tub.pip.shal.well.y | . | . |
| dug.well.y | . | . |
| protec.spring.y | . | . |
| unprotec.spring.y | -0.0327 | 0.9679 |
| lake.riv.rain.y | 0.0000 | 1.0000 |
| peddler.y | . | . |
| bottled.wat.y | -0.0158 | 0.9843 |
| others.y | . | . |
| pop.dens.km2 | 0.0003 | 1.0003 |

Table 7: Coefficients Elastic Net Poisson Regression with $\lambda = \lambda_{1SE}$.

| Variables | Neuron 1 | Neuron 2 |
|---|---|---|
| weight | 3.090311 | 2.709838 |
| bias | 0.780622 | -0.23502 |
| pop.nr.male | 0.519813 | 0.083543 |
| pop.nr.female | -0.46333 | 0.052662 |
| rain.val.mm | 0.04555 | 3.915465 |
| rain.int.mm.hr | -0.01621 | 2.843544 |
| wind.spe.kph | 0.084411 | -1.3005 |
| wind.spe.max | -0.02869 | -0.12245 |
| air.tem. | -0.03311 | 0.135358 |
| air.tem.max | 0.037388 | 2.526867 |
| air.tem.min | 0.060432 | 0.336202 |
| air.pre | -0.02916 | 1.508606 |
| dew.poi | -0.01533 | -0.33436 |
| relative.hum | -0.00738 | 2.12003 |
| relative.hum.max | 0.011397 | -1.41103 |
| relative.hum.min | 0.006269 | -0.75145 |
| sunshine.dur.hour | 0.008773 | -0.94164 |
| solar.rad.w.m2 | -0.02311 | 0.908761 |
| soil.moi.vwc | -0.01418 | -0.57902 |
| soil.moi.cm | 0.023234 | -1.54118 |
| soil.tem.mean | -0.00189 | 1.665815 |
| soil.tem.mean2 | -5.86E-05 | -0.81049 |
| carabao | -0.00933 | 1.433683 |
| cattle | 0.043424 | -4.09695 |
| Goat | -0.00731 | 5.079513 |
| Hog | -0.01182 | -0.1356 |
| Number.of.Households | 0.00071 | 0.238961 |
| own.fauc.com.wat.sys | -0.00378 | -0.13511 |
| shar.fauc.com.wat.sys | 0.019145 | -0.64881 |
| own.tub.pip.dep.well | -0.02127 | 1.221236 |
| shar.tub.pip.dep.well | 0.011098 | -0.42015 |
| tub.pip.shal.well | -0.00166 | 0.733523 |
| dug.well | 0.004039 | -0.80716 |
| protec.spring | -0.03715 | -2.44097 |
| unprotec.spring | 0.021065 | -1.6285 |
| lake.riv.rain | -0.01229 | -0.76821 |
| peddler | -0.00337 | 2.770905 |
| bottled.wat | 0.014114 | -0.73865 |
| others | -0.01311 | 2.455004 |
| pop.dens.km2 | 0.018748 | 1.130989 |

Table 8: Coefficients Bayesian Neural Network.