LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
School of Business and Management
Degree Program in Computer Science

SAINT PETERSBURG NATIONAL RESEARCH UNIVERSITY OF INFORMATION TECHNOLOGIES MECHANICS AND OPTICS (ITMO UNIVERSITY)
Software Development Chair Faculty of Infocommunication Technologies
Master's Programme in Software in Infocommunications

Andrei Gorbulin

Missing data analysis in emotion recognition

1st Supervisor/Examiner: Prof. Ajantha Dahanayake, PhD, LUT

2nd Supervisor/Examiner: Assoc. Prof., Nikita Osipov, Candidate of Engineering sciences, ITMO University

Lappeenranta – Saint Petersburg
2018

ABSTRACT

Author: Andrei Gorbulin

Title: Missing data analysis in emotion recognition

Department: LUT School of Business and Management, Innovation and Software

ITMO University, Software Development Chair, Faculty of Infocommunication Technologies

Master's Programme:

Double Degree Programme between LUT Computer Science and ITMO Software in Infocommunications

Year: 2018

Master's thesis: Lappeenranta University of Technology
ITMO University
51 pages, 7 tables, 26 figures, 1 appendix

Examiners: Prof. Ajantha Dahanayake, PhD, LUT
Assoc. Prof., Nikita Osipov, Candidate of Engineering sciences, ITMO University

Keywords: missing data, emotion recognition, EEG, listwise deletion, pairwise deletion, hot deck imputation, linear regression analysis, EM algorithm

Missing data is a widespread fundamental problem that cannot be ignored. It distorts the data, sometimes even to the point where it is impossible to analyze data at all. In emotion recognition, it was discovered that one of the best approaches to identify human emotions is by analyzing EEG (electroencephalography) results combined with peripheral signals. In this thesis EEG data is used to test which missing data techniques are more efficient and reliable in emotion recognition. During the research, the software was created, which implicates all the methods that are tested. In the end, author concludes which techniques should be used in emotion recognition and when.

РЕЗЮМЕ

Автор: Горбулин Андрей Сергеевич

Заглавие: Анализ данных в условиях неопределенности исходной информации, полученных при распознавании экспрессии человека

Факультет: ЛТУ Факультет Бизнеса и Менеджмента

Университет ИТМО Кафедра Программных Систем Факультет Инфокоммуникационных Технологий

Магистратура: Программное обеспечение в инфокоммуникациях

Год: 2018

Диссертация: Лаппеенрантский Технологический Университет,
Университет ИТМО,
51 страница, 7 таблиц, 26 рисунков, 1 приложение

Экзаменаторы: Профессор Аджанта Даханайаке
Доцент, к.т.н. Осипов Никита Алексеевич

Ключевые слова: пропуски в массивах данных, распознавание эмоций, распознавание экспрессии, анализ полных наблюдений, заполнение с подбором

Проблема наличия пропусков в массивах данных является фундаментальной проблемой. Она широко распространена в различных сферах деятельности и не может быть игнорирована. Эта проблема приводит к искажению данных, иногда до такой степени, что становится невозможным проводить какой-либо их дальнейший анализ. Учеными было обнаружено, что одним из наилучших способов распознать человеческие эмоции является анализ результатов ЭЭГ (электроэнцефалографии) вместе с периферийными сигналами. В этой работе данные ЭЭГ используются для тестирования различных методов борьбы с пропусками, чтобы определить какие из них являются эффективными и надежными. В ходе исследования было создано программное обеспечение, которое реализует эти методы. В конце диссертации автор делает выводы о том, какие из методов должны использоваться в борьбе с пропущенными данными в распознавании эмоций и когда.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

OAR	Observed at random
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
EEG	Electroencephalography
MEG	Magnetoencephalography

1. INTRODUCTION

People use emotions to communicate with each other on a daily basis. Moreover, even a simple conversation between two people involves emotions to convey the message. This mechanism allows people to understand each other better and behave according to the situation, because the same sentence can be treated differently depending on emotions person who says it has.

Due to the emotions being so important in human to human interaction, it is essential that interaction between human and machine is also based on emotions [1]. Furthermore, it has been proven that human would feel more comfortably engaging with machines that can react to their emotions [2].

To be able to continue, it is essential that certain terms are defined to avoid ambiguity:

- Emotion recognition is a technique that allows machines to detect and correctly recognize human emotions;
- A facial expression is a visible manifestation of the affective state, cognitive activity, intention, personality, and psychopathology of a person [3].

Emotions can be expressed both verbally and non-verbally. It has been discovered that it is better to use non-verbal methods, because they yield more reliable information [1]. Using non-verbal methods, information can be collected differently: by analyzing gestures, facial expressions or even using electroencephalography or magnetoencephalography [4]. An electroencephalography (EEG) is a test that detects electrical activity in brain using small, flat metal discs (electrodes) attached to scalp [5]. Magnetoencephalography (MEG) a noninvasive technique that detects and records the magnetic field associated with electrical activity in the brain [6]. Detailed explanation and principles of EEG and MEG can be found in respective articles [7, 8].

By being able to recognize emotions correctly we can make existing technologies more human-friendly as well as create new ones. For example, by using emotion recognition,

nursing robots can have smooth user interaction, which is especially important in this area [9]. Smart cities can also use this technology in order to have information about its inhabitants' behavior. This will reduce crime rates and prevent several mental problems people might have by notifying respective specialists in time. Additionally, this feature can be used in Virtual Reality teaching [10]: by understanding what students can feel during the lecture, teacher will have some feedback about his work, which will give him an opportunity to improve.

Despite its usefulness, emotion recognition has some major challenges and problems. In this thesis, we will focus on missing data.

1.1. The research problem

Missing data are observations which are planned and are missing [11]. This is a fundamental problem that can be stumbled upon during any experiments or surveys. Furthermore, any kind of data acquisition is always in danger of getting results with missing values. This is by no means exception to emotion recognition. Actually, getting missing data results is one of the main problems in recognizing emotions [1, 12, 13], particularly in face recognition [14-26].

For example, during data acquisition process in EEG, some data is inevitably missed due to power line interference, motion artifacts, electrode contact noise and sensor device failure [1]. Consequently, missing data distorts the results, which may lead to incorrect conclusions regarding the emotion that is displayed. Nevertheless, this problem was completely ignored for decades and only recently started to get addressed [12].

If emotion recognition techniques will be used in the future, it is extremely important that human emotions are recognized correctly. For nursing robots, recognized emotions are deciding factor in how they would approach interaction with a person. In emotion recognition, handling missing data correctly will give researchers working on this technology more precise data, which will increase percentage of correctly recognized emotions.

1.2. Research objectives and questions

Considering the research problem, there are three research questions:

RQ1. What are the techniques that can be used to deal with missing data?
RQ2. Which techniques are the most suitable to use in emotion recognition?
RQ3. When each of those suitable techniques can be used?

The research tasks and objectives are:

1. Definition and classification of missing data.
2. Classification of techniques that are used to deal with missing data.
3. Comparison of techniques and their uses regarding emotion recognition.
4. Conclusion regarding use of different techniques in emotion recognition.

1.3. Research methodology

This research consists of three major steps:

1. Acquiring theoretical background on missing data;
2. Studying missing data in relation to emotion recognition;
3. Implementing software to help with missing data problems in emotion recognition, testing it on real data.

Based on these steps, detailed project plan that is used in this thesis is established:

1. Defining missing data;
2. Overviewing techniques that are used to address this problem;
3. Studying missing data in relation to emotion recognition;
4. Engineering software that will apply these techniques;
5. Testing software on a real data;
6. Concluding the results, finding the best technique.

1.4. Phases of the research

There are total of 9 phases of research. Deadlines and results after each phase are concluded in Table 1.

Table 1. Phases of the research.

Phase	Result	Deadline
Topic selection.	Thesis topic is selected.	30.10.2017
Research question formulation.	Research questions are concluded.	08.01.2018
Literature review.	Literature analysis is made, relevant articles are selected.	25.01.2018
Writing a proposal.	Proposal is submitted.	01.02.2018
Post-mortem of proposal.	Proposal drawbacks are taken into consideration.	02.02.2018
Classification of missing data and techniques to deal with it.	Chapters about missing data and techniques to deal with it is completed.	30.02.2018
Comparison of techniques and their uses regarding emotion recognition.	Different techniques are compared, most suitable techniques are chosen for emotion recognition.	30.03.2018
Finalization of a thesis.	Thesis is completed.	15.04.2018
Preparing presentation.	Presentation for thesis is ready.	27.04.2018

1.5. Resources required

Hardware and software

Personal computer and necessary software for the research.

Time

Time is strictly limited; research must be finished by 27.04.2018.

Relevant articles

Due to research being based on other publications, relevant articles should also be considered as a resource.

1.6. Outcomes of the research

This research project has following outcomes by the end of the project:

1. Different techniques for dealing with missing data are compared regarding their use in emotion recognition;
2. The best technique to be used is proposed for different cases.

2. SYSTEMATIC LITERATURE REVIEW APPROACH

Several databases are used to search for relevant articles, such as: IEEE, ACM, Springer, Web of Knowledge and LUT Finna.

The keywords used for the search are: “missing data”, “face recognition”, “emotion recognition”, “missing data” AND “face recognition”, “missing data” AND “emotion recognition”.

The search has been made for the articles using mentioned databases. Firstly, search has been made for the key phrase itself, then the results are narrowed down by using AND operation. Tables 2 and 3 provide information about the number of articles on each key phrase in 5 different databases.

Table 2. Number of articles in databases

Database name	**“missing data”**	**“face recognition”**	**“emotion recognition”**
IEEE	2097	21476	5312
ACM	249	432	401
Springer	90457	17787	5163
Web of Knowledge	19037	21973	6780
LUT Finna	185293	26323	9769

Table 3. Number of articles in databases

Database name	**“missing data” AND “face recognition”**	**“missing data” AND “emotion recognition”**
IEEE	30	6
ACM	0	0
Springer	362	161
Web of Knowledge	35	3
LUT Finna	488	332

It is important to mention that these tables only provide information about the quantity of articles. Some of these articles are irrelevant to the question, so only relevant articles have been chosen. If needed, it is possible to revisit these keywords to get more articles.

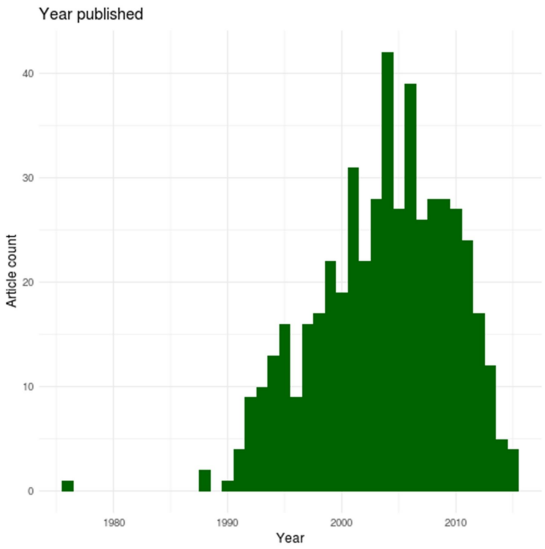

Figure 1. Publication years for "missing data" keyword

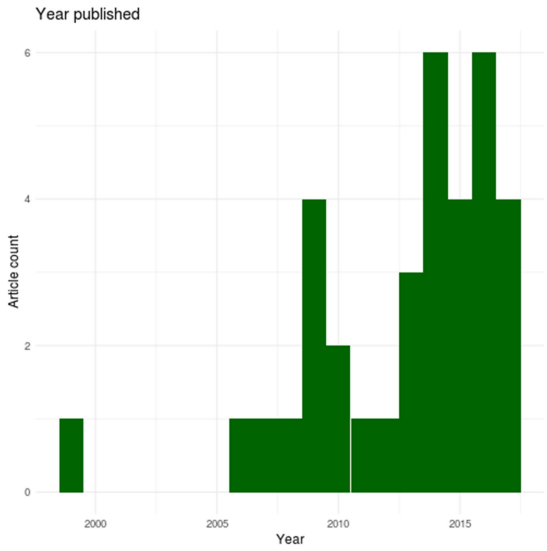

Figure 2. Publication years for "missing data" AND "face recognition" keyword

Articles from Web of Science are analyzed using NAILS tool [27] to get an overview of field's status. Analysis is shown in Figures 1 and 2. It is clear that even though missing data is popular around 2005 and lost its popularity, missing data in face recognition is gaining attention more and more in last 4 years.

3. OVERVIEW OF MISSING DATA

3.1. Classification of missing data

In the field of emotion (facial) recognition the missing data problem occurs quite often [1, 12, 13]. This problem is widespread and occurs not only in this particular field, but also in sociology [28], political science [29], psychology [30], education [31] and communication [32]. Traditionally, reasons that lead to partial absence of data are impossibility of obtaining or processing data, distortion or intentional hiding of information. Consequently, incomplete data is thrown into input of programs that analyze data.

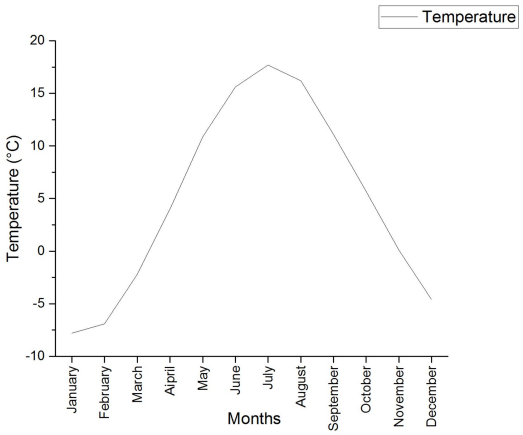

Figure 3. Average temperature in Saint Petersburg throughout the year

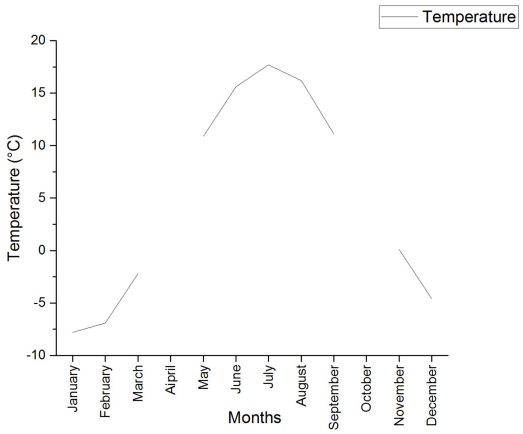

Figure 4. An example of missing data while measuring temperature

Let's take a look on an example of missing data. Complete data about average temperature in Saint Petersburg [33] is shown in Figure 3. Now, let's assume that during April and October measurement equipment is broken and we couldn't measure temperature during these periods of time. This means that now we have to deal with missing values in the middle of our measurement. An example of this is an illustration shown in Figure 4. As you can see, while the measurement is conducted, data for April and October went missing and we still have to put this data into analytic software, which leads to a problem.

So, how do we deal with this? To answer this question, we need to first determine if all missing data cases should be treated equally or not. And, as it turns out, missing values can have different nature and can generally be split into three groups [34]. Let's go ahead and determine those groups first.

It is fairly obvious, that only 2 cases are possible while using multiple variables [35]:

1. Missing data problem occurs only in one variable;
2. Missing data problem occurs in multiple variables.

For the sake of simplicity, we will only look at first case. Let's assume that we only have 2 variables – X and Y. Let X be the variable that has all data in it, then Y will be the variable containing the missing data. You can see an example of missing data monotone pattern in Figure 5: here X is a complete variable, while Y is incomplete.

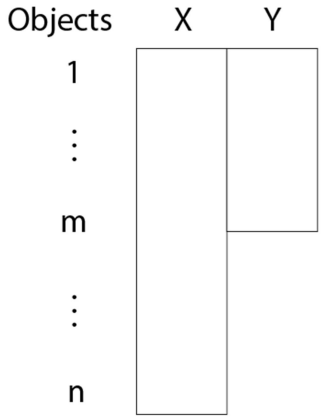

Figure 5. An example of monotone pattern containing missing data in one variable [35]

As Rubin and his colleagues suggest [34, 35], it is incredibly useful to classify the missing data mechanism according to whether probability of response:

1. Depends on Y, and, possibly, on X;
2. Depends on X, but not on Y;
3. Independent from both X and Y.

In third case we can say that missing data is missing at random (MAR) and observed data is observed at random (OAR), thus data is missing completely at random (MCAR). This condition is very strict, so in reality data can pretty much never be considered MCAR due to the nature of experiments [36, 37].

In second case we can say that missing data is missing at random (MAR). It means that data is missing equally not throughout the entire variable, but inside groups of variables. For example, if during a sociological research missing data is more likely to occur in men's responses than women's, but inside those 2 groups will occur at random, then this data is MAR.

In first case, data is neither MAR or OAR. In this case we can say that data is missing not at random (MNAR). As a rule, reason behind missing data in this case is hidden in the field of study itself.

Additional information about missing data mechanisms can be found in related materials [34, 35, 38-40].

3.2. Techniques for dealing with missing data

While applying different techniques for dealing with missing data, it is important to know what kind of missing data you encountered: MCAR, MAR or MNAR [35]. If data is complying to MAR condition or MCAR condition as more strict, then missing data mechanism is ignorable and different techniques of dealing with missing data can be applied. However, if data is not complying with these conditions (is considered MNAR), then missing data mechanism is non-ignorable and knowledge of this mechanism is essential for correct analysis of data. Moreover, MCAR and MAR provides unbiased parameter estimates, while MNAR gives us biased parameter estimates [41].

You can build tests that distinct between MCAR and MAR [42], but you can't always recognize if missing data mechanism is ignorable or not.

So, what are the techniques that can be used to deal with missing data problem? There are a lot of techniques that can be used for this purpose [40]. However, it would be useful to split them into two separate groups:

1. Available data analysis;
2. Using imputation methods.

Correspondingly, as shown on Figure 6, all techniques can be split into two groups.

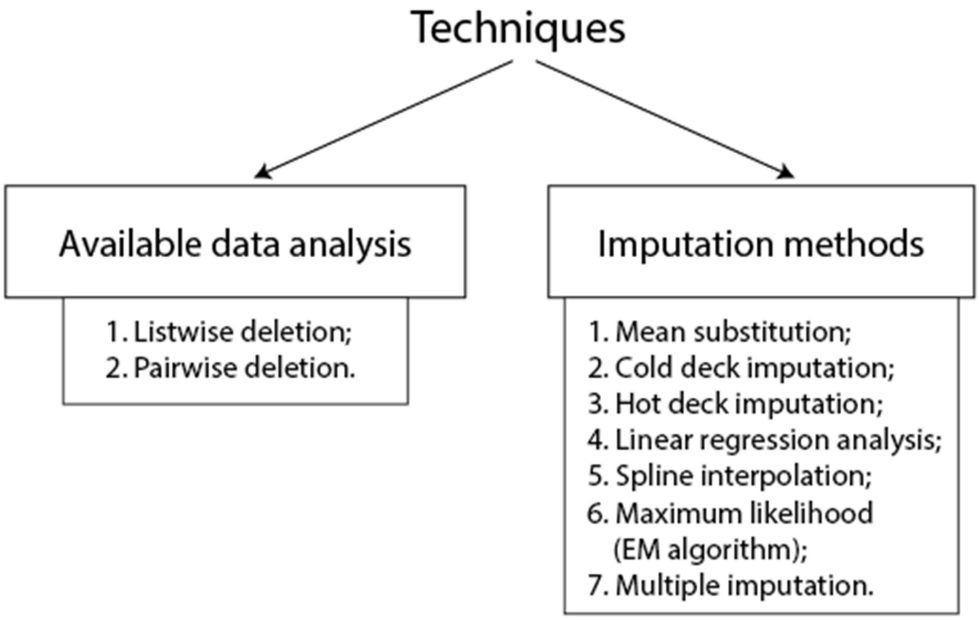

Figure 6. Two groups of techniques for dealing with missing data

As you might have noticed, two of imputation methods are missing on this list. They are Bartlett's method [35] and ZET algorithm [43-45]. These methods are rarely used compare to all the other ones listed above. As a consequence, they are removed from this research due to their high complexity and low frequency of use. However, in future works, they should be included in experiments as all the other methods are.

Now, let's dive in and take a closer look at each of these groups.

4. AVAILABLE DATA ANALYSIS TECHNIQUES

There are two techniques that can be marked out in this group of methods:

1. Listwise deletion;
2. Pairwise deletion.

It is necessary that MCAR condition is met to use these methods. Quite often this becomes a big problem due to the fact that usually data is complying to MAR conditions, but not MCAR [36, 37]. Also, amount of absent data should be relatively small in order to use these methods, otherwise they will lead to biased parameter estimates, though bias can be considered minimal [46, 47].

Methods that are included in “Available data analysis” group are shown in Figure 7.

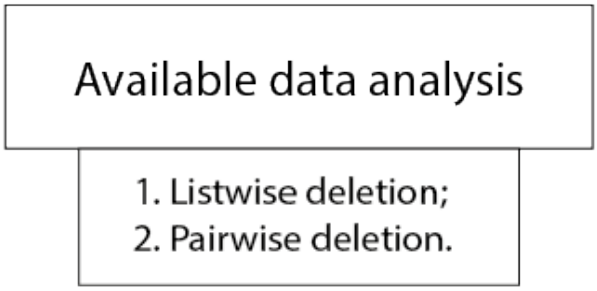

Figure 7. Methods that are included in “Available data analysis” group

4.1. Listwise deletion

Listwise deletion (also known as complete case analysis) [40, 41, 48] is a method that comes to mind first when you think about missing data problem. The whole point of this method is to remove from data set all the results that have missing values in them. It does not require any restoration of data, only conducting following analysis without incomplete objects.

Consider you have hypothetical data that has X as complete variable and Y as incomplete variable. It is easy to see how this method will affect data from illustrations shown on Figures

8 and 9: graph of full data and graph of data that remained after listwise deletion. Black dots represent complete data and red dots represent objects that have missing information about Y variable.

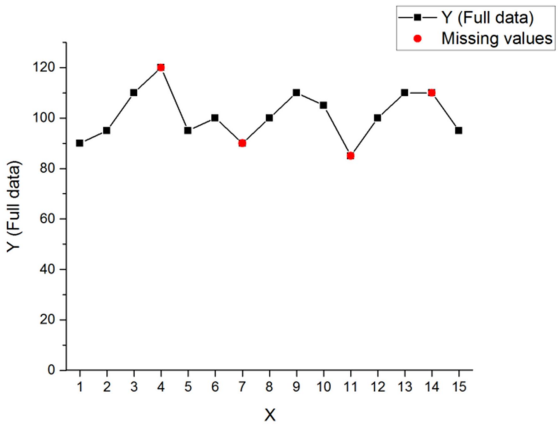

Figure 8. Full data before listwise deletion

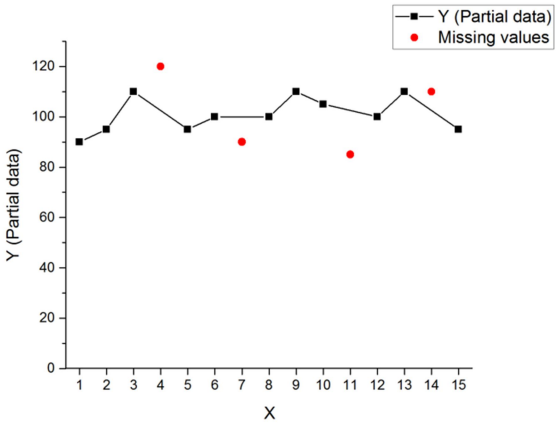

Figure 9. Data set after listwise deletion

This method is used by default by most researchers [49], although in reality it is "a method that is known to be one of the worst available" [32]. According to [49], this method should be given up in favor of hot-deck imputation method (see Section 5.3).

4.2. Pairwise deletion

Pairwise deletion (also known as Available case analysis) [40, 41, 48] is very similar to a previous one, except when having missing values in several variables, other variables would have counted in following analysis. This way, if respondent did not answer all the questions, his participation will not be ignored, unlike first case; questions that he answered will be analyzed, but questions he did not answer will not.

The hypothetical data example is shown in Table 4. Here X is a complete variable, while Y, Z and T are non-complete variables. In this case, unlike listwise deletion, observations *j* and *k* will not be completely ignored: non-complete variables Z and T of observation *j* will still contribute towards future analysis as well as T variable of observation *k*.

Table 4. Pairwise deletion example (hypothetical data)

Variable	Observation *i*	Observation *j*	Observation *k*
X	54	35	25
Y	67	?	?
Z	75	48	?
T	58	98	26

Even though this method can get better results depending on circumstances, it still has the same problems as listwise deletion [35, 40].

5. IMPUTATION METHODS

Imputation methods are alternative methods to deal with missing data. The main idea of this group of methods is that instead of removing some entries in data in order to cope with missing data, new values will be assigned to them. This way, all the missing values will be filled in, so data set will look complete and can be analyzed further.

Generally speaking, these methods provide better results compare to previous ones due to the reasons stated above [35, 40]. Nevertheless, imputation techniques require more computing power, which may be essential in some cases.

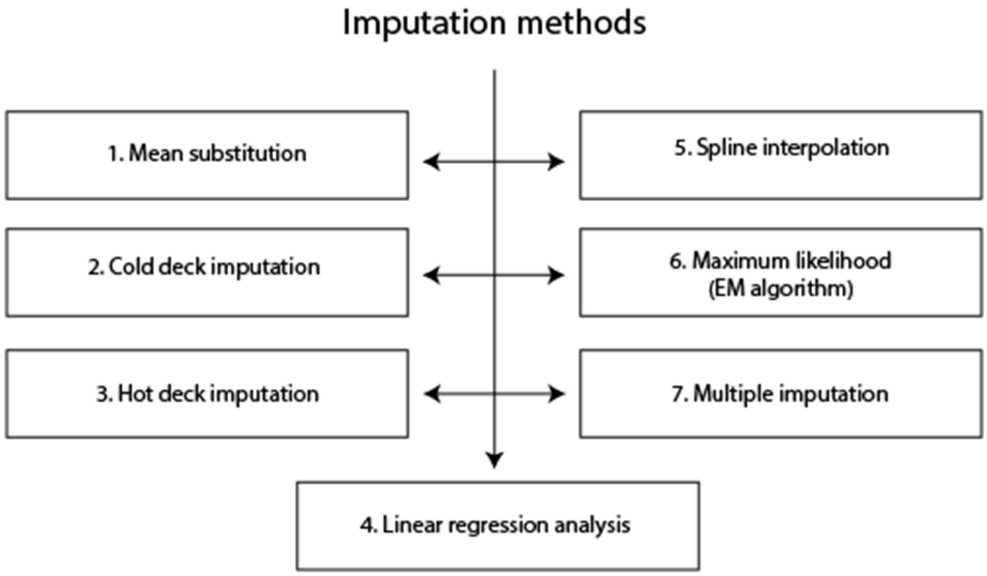

Figure 10. Imputation methods

This group of methods contains different techniques, such as:

1. Mean substitution;
2. Cold deck imputation;
3. Hot deck imputation;
4. Regression analysis;

5. Spline interpolation;
6. Maximum likelihood (EM algorithm);
7. Multiple imputation method.

Figure 10 represents all the imputation methods that are discussed in this thesis.

5.1. Mean substitution

Mean substitution [40, 41, 48] is a wide-spread technique used by many researchers [50]. The core idea is that you simply replace missing values with mean of the values observed. In Figure 11 you can see an example of mean substitution on hypothetical data. Black dots here represent observed values; red dots represent missing values that have been replaced with mean.

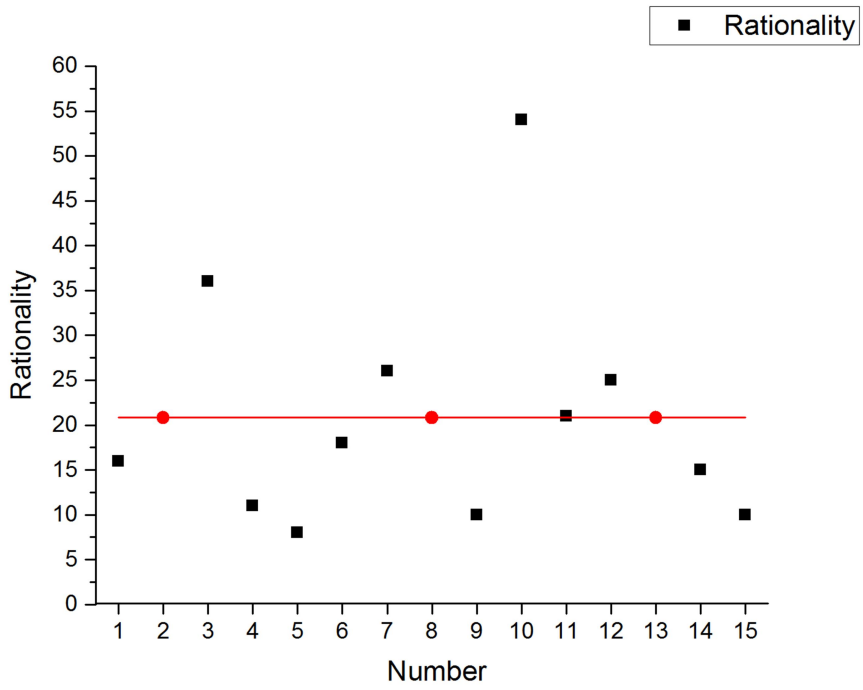

Figure 11. Mean substitution technique (hypothetical data). Mean value is around 21.

This technique has a lot of problems. First of all, if there is a lot of missing values, it would be incorrect to replace them with mean value: it would lower the dispersion significantly [49].

Also, considering the fact that quite often missing values can be much lower (or higher) than mean value, this approach should only been used in some specific cases and shouldn't be go-to strategy for handling missing values [51].

5.2. Cold deck imputation

In this method, missing data found in input data is simply imputed with static external value (usually 0) [52]. Quite often this value is based on the previous researches that have been done on the same topic.

There is not much to be said on this method: it provides false data just to get rid of missing values, but it usually is better to just use available data analysis methods in this case.

5.3. Hot deck imputation

Main idea of this method is to substitute missing value with value of another observation that is the closest to this one [49]. The thinking process is as follows: if observation is close to another observation, then it should take all the similar parameters.

Take a look at an example shown in Table 5. Here we have 2 observations: observation *i* and observation *j*. When we encounter observation *k*, that has Z variable missing we compare *k* to both *i* and *j* using X and Y. In this example, *k* is much closer to *i* than to *j*, which means that Z should be taken from there, making it 67.

Table 5. Hot deck imputation example (hypothetical data)

Variable	Observation *i*	Observation *j*	Observation *k*	Observation *k* (restored)
X	53	25	45	45
Y	85	76	84	84
Z	67	23	?	67

Some researchers seriously advise to use this method instead of listwise and pairwise deletions as well as mean substitution [49]. This is mainly due to the fact that according to Roth [30], this method "can be both valid and simultaneously easy to use".

This method also has a significant downside: a consistent theory is still not well developed on when and where to use this method, and what effect it has on statistical properties of data [53].

5.4. Linear regression analysis

Linear regression analysis [54] is based on the assumption that variables are collinear (somehow distributed around the linear function). So, values of this linear function can be used to substitute the missing values.

As Draper mentioned [54], if you consider data lying on the linear function, then this function can be calculated as follows (for 2 variables case):

$$y(x) = a + bx.$$

The coefficients *a* and *b* can be acquired using following formulas:

$$b = \frac{n * \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{n * \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} \text{ and}$$

$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i}{n},$$

where *n* – amount of observed data values.

The example of this method is shown on Figure 12. As you can see, all missing values have been substituted with values that are lying on the line, which is determined with above formulas.

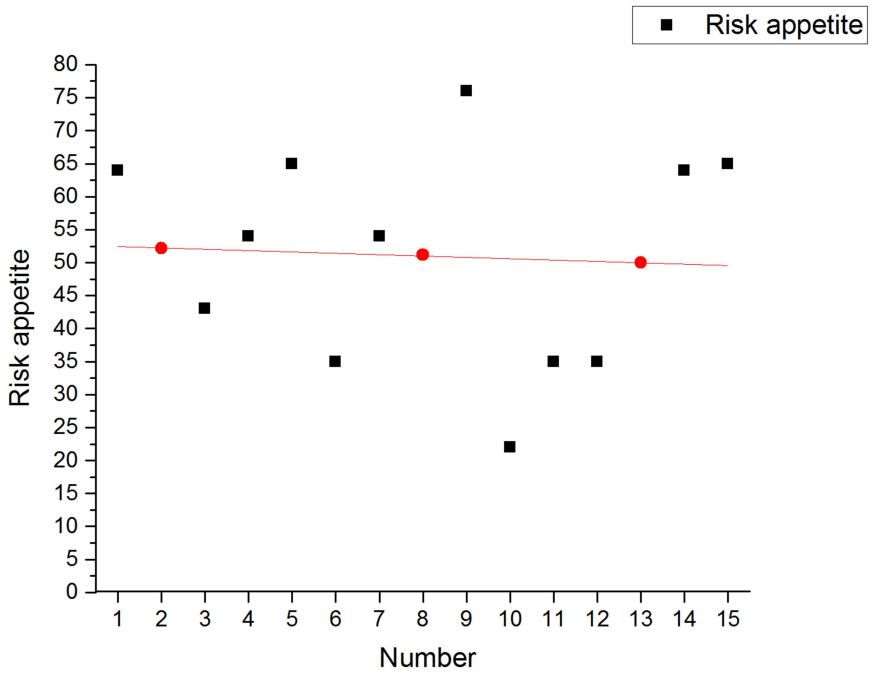

Figure 12. Substitution of missing values using linear regression method (hypothetical data).

The main advantage of this method is that theory is well developed [55], however method also has a major downside. In order to use this method, you have to prove that variables are collinear, and in some areas collinearity problem has been recognized as a serious problem [54, 56, 57]. More information about collinearity problem in regression analysis can be found in paper [58].

5.5. Spline interpolation

Spline interpolation method is a method that uses splines to predict missing values [59]. Firstly, you create a spline function using available discrete values. After that, you just use values of recreated function to substitute missing values. It is important to know that some splines like B-splines do not necessarily have to have observed values; but nevertheless, they use their weighted sum as a reference to create a function.

In this thesis, we will stick to use only B-splines, which can be acquired using Cox-de Boor's algorithm [59, 60]:

$$y(x) = \sum_{i=0}^{n} N_i^k(x) * y_i,$$

where k is an order of spline, $N_i^k(x)$ is a function, described as:

$$N_i^1(x) = \begin{cases} 1 \text{ if } x_i \leq x \leq x_{i+1} \\ 0 \quad \text{otherwise} \end{cases},$$

$$N_i^k(x) = \frac{x - x_i}{x_{i+k-1} - x_i} N_i^{k-1}(x) + \frac{x_{i+k} - x}{x_{i+k} - x_{i+1}} N_{i+1}^{k-1}(x).$$

Main disadvantage of this method is that if you have a lot of missing values, then precision of recreated values will be very low, which means that recreated values will be very different from the ones that could be observed [59].

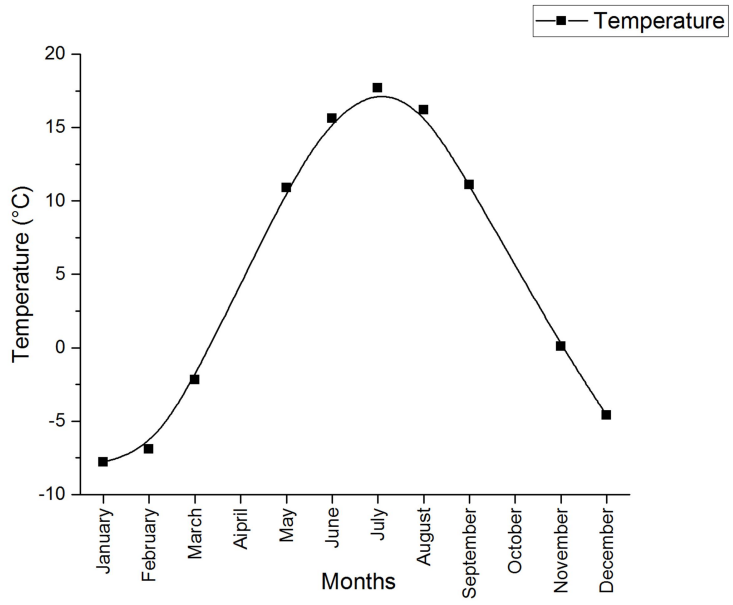

Figure 13. Using spline interpolation method to acquire missing values

Figure 13 illustrates weather in Saint Petersburg throughout a year provided by [33]. The data for April and October is missing, so spline interpolation method was used to recreate it. Using observed values a function was created, which can be used to fill in the missing data.

5.6. Maximum likelihood (EM algorithm)

This method is proposed by Rubin and his colleagues [61] in 1977. EM algorithm is called so because it consists of 2 major steps:

1. Expectation step;
2. Maximization step.

The core idea is to create a likelihood function of a statistical data provided by given observations and then find local maximums of this function. Detailed mathematical explanation of this method can be found here [61, 62]. The method is explained using an example provided in [63] (all pictures are taken from there as well).

Let's consider that we have a hypothetical data set that has 2 different groups represented by red and blue dots (Figure 14). We can easily calculate mean and other parameters that can characterize these groups. For example, mean value of a red group is around 3 and mean value of a blue group is around 7.

As it turns out, using EM algorithm we do not actually need to know the colors of dots to calculate its statistical parameters (Figure 15).

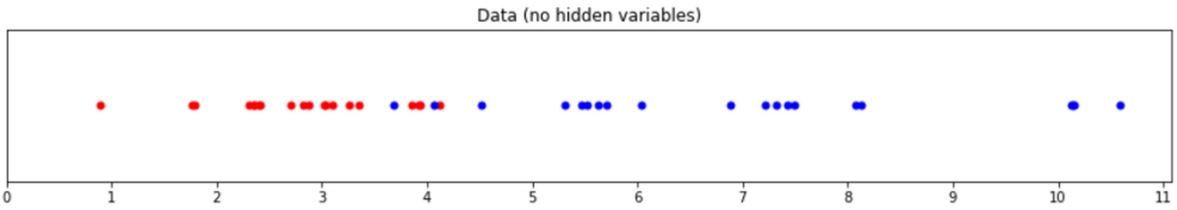

Figure 14. Hypothetical data set with 2 distinct groups of observed values [63]

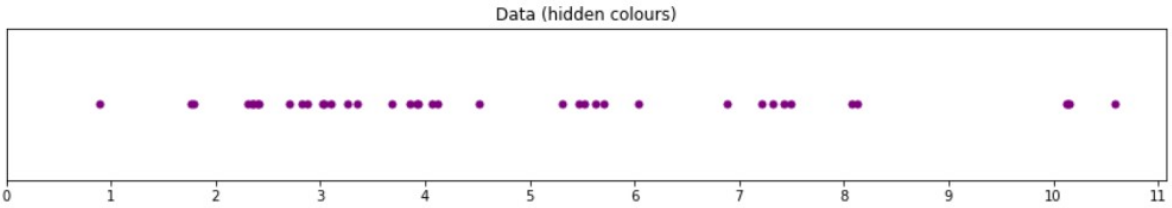

Figure 15. Same data set, but groups are hidden [63]

In order to do this, we should proceed following steps:

1. Make a blind guess on initial value on the parameter we are looking for;
2. E-step: Compute the likelihood that each parameter produces data point;
3. For each data point we calculate weights that show the likelihood of the parameter producing those points;
4. M-step: Calculate a better value for the parameter using maximization of likelihood function;
5. Repeat steps 2-4 until parameter estimate converges or required number of steps is completed.

The major problem is we never know how many groups of dots we have, so we have to make a guess here.

A walkthrough of all the steps of the example:

1. We have to make blind guesses here. Let us assume that we have 2 groups of dots with mean values of 1.7 and 9. This assumption gives us the distribution shown in Figure 16.
2. Now for each dot we calculate numbers that will represent probabilities of being in each group. In this case, with our current guesses the data point at 1.761 is much more likely to be red (0.189) than blue (0.00003).

3. Now we turn those numbers into actual probabilities by calculating their weights. For data point at 1.761 it would give us 99.98% chance of being in red group and around 0.02% chance of being in a blue group.
4. Now we calculate more suitable values for the mean parameters that we blindly guessed at step 1.
5. Jump to step 2 and do a new iteration.

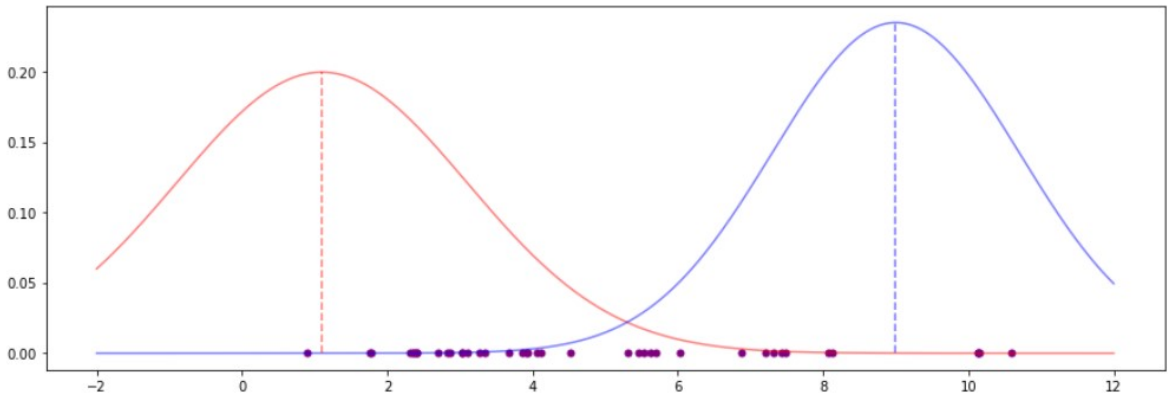

Figure 16. Likelihood function provided by our blind guess [63]

So, by calculating more and more iterations, we get better and better values for the parameter that we are looking for. Figure 17 shows first 5 iterations for our example (recent iterations have stronger appearance).

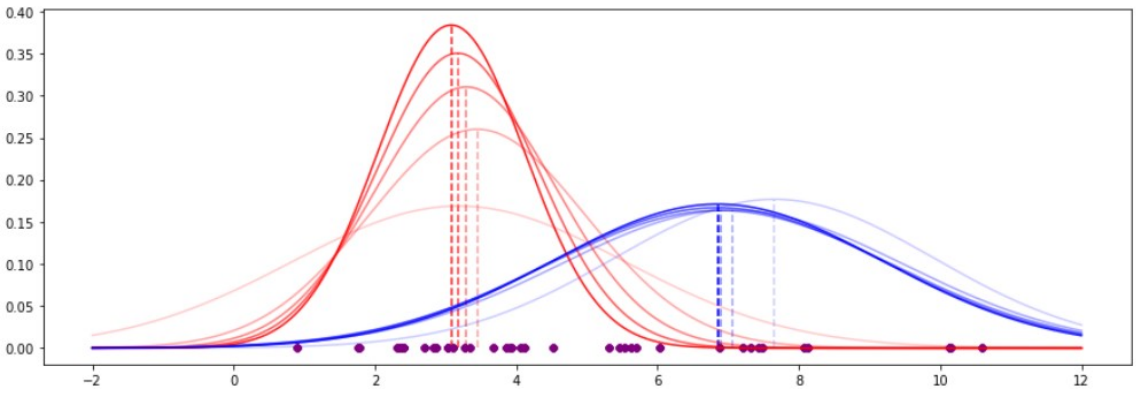

Figure 17. Likelihood functions (recent ones have stronger appearance) [63]

And, finally, after 20 iterations we will have a picture shown on Figure 18 that will give us reasonably close mean values for both groups. Calculated mean values would give us 2.91 for red and 6.84 for blue dots, while real values are 2.8 and 6.93, respectively.

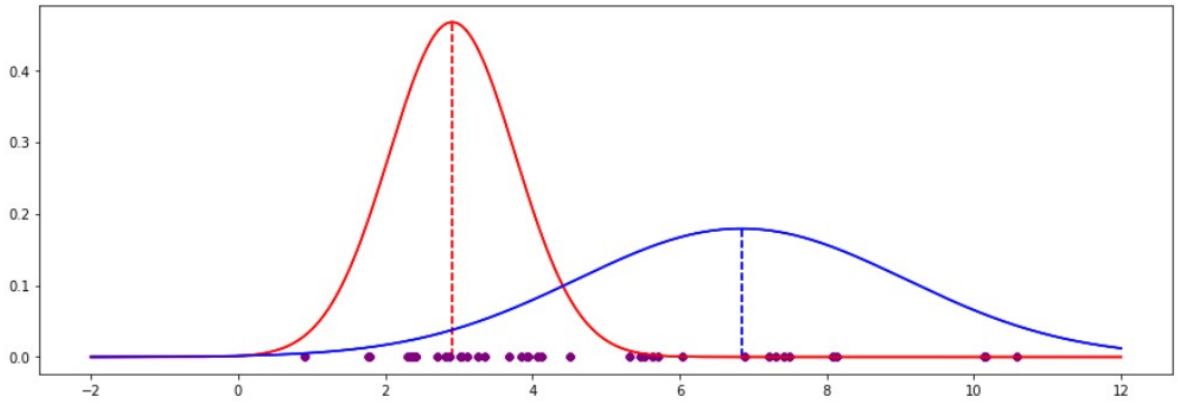

Figure 18. Likelihood function after 20th iteration [63]

5.7. Multiple imputation

All previous imputation methods are also known as single imputation methods. This means that you replace missing value with one other value that is calculated by the algorithm.

On the other hand, multiple imputation method [35] replaces 1 missing value with multiple values complying with different models. Only a brief explanation of this method would be given in this thesis; additional information can be found in related sources [35, 64].

So, how does this method work? From Figure 19 you can see that one missing value is represented by m different values. This way, if you have m models that can be used to impute values, at the end you will receive m new data sets with imputed values. In order to perform tests, you need to perform them on all possible values, which mean if you have n number of observations that you have to run the tests n^m times. Simplified schema of multiple imputation method is given on Figure 20.

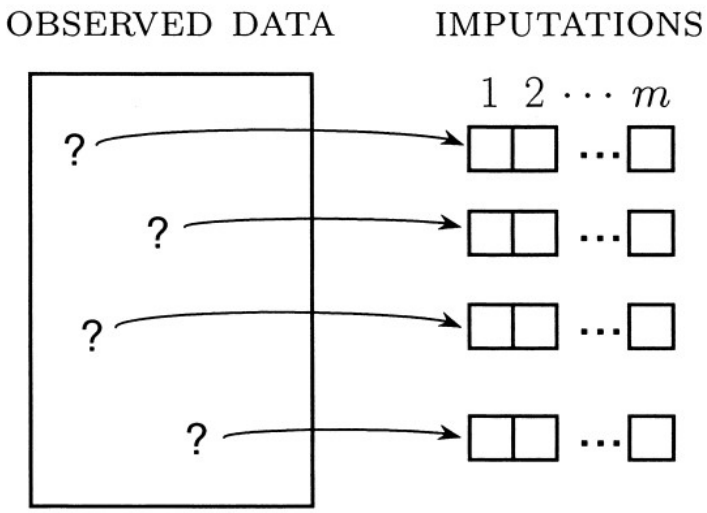

Figure 19. Multiple imputation method [40]

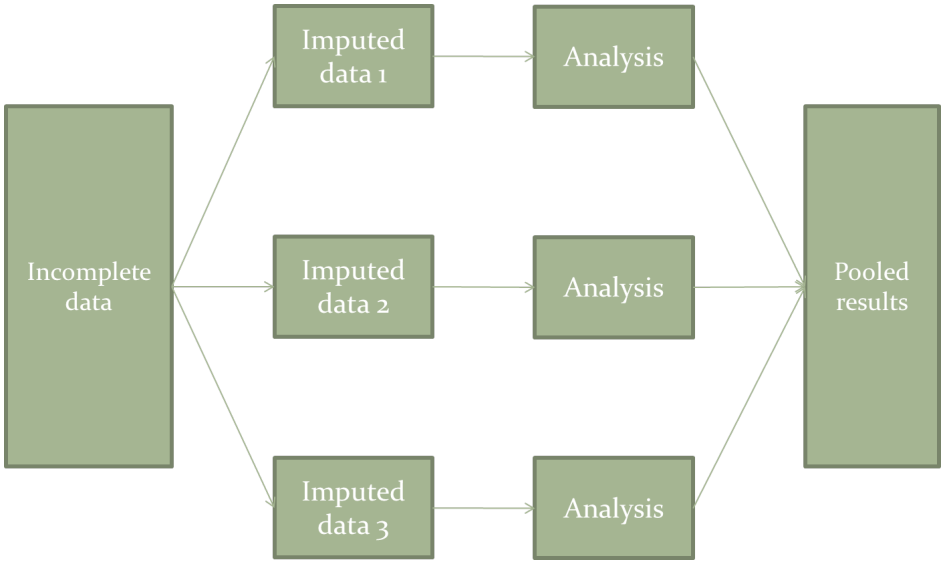

Figure 20. Schema of multiple imputation method

When implementing this method, extra caution should be given, because "…a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests." [65]. This is mainly due to its high demand on computer's resources.

6. MISSING DATA IN EMOTION RECOGNITION

Modern emotion recognition researches are based on circumplex model proposed by Russel in 1970 [66]. Particularly, this model suggests that all human emotions can be found in two-dimensional space. The successor to this model is Arousal-valence model (see Figure 21). This way, to determine person's emotion it is enough to know his arousal and valence. And this is exactly what modern researchers do [1, 12, 67].

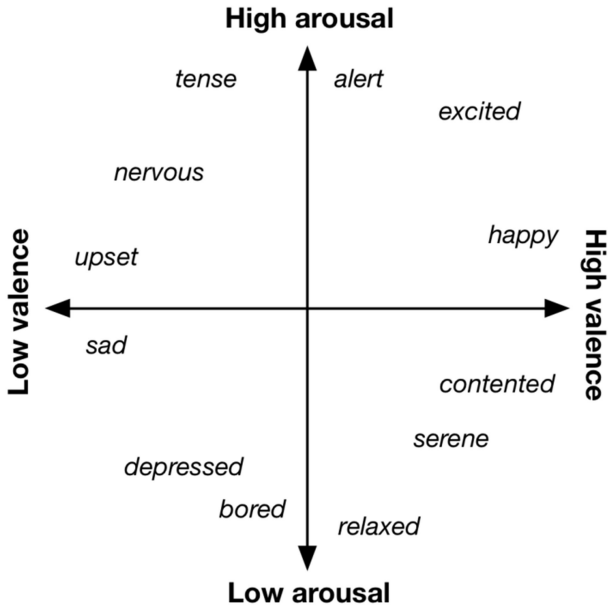

Figure 21. Arousal-valence model [68]

As it was already mentioned in introduction, missing data problem in emotion recognition was completely ignored until 2011 [12]. In his study, Wagner uses several methods to deal with missing data, such as: listwise deletion, pairwise deletion and cold deck imputation.

Since then, only 2 studies addressed missing data problem in emotion recognition. One of these studies [1] used restricted Boltzmann machine, which is a maximum likelihood algorithm. As this study showed, their approach has much better results than methods used in previous study [12]. You can see related picture in Figure 22.

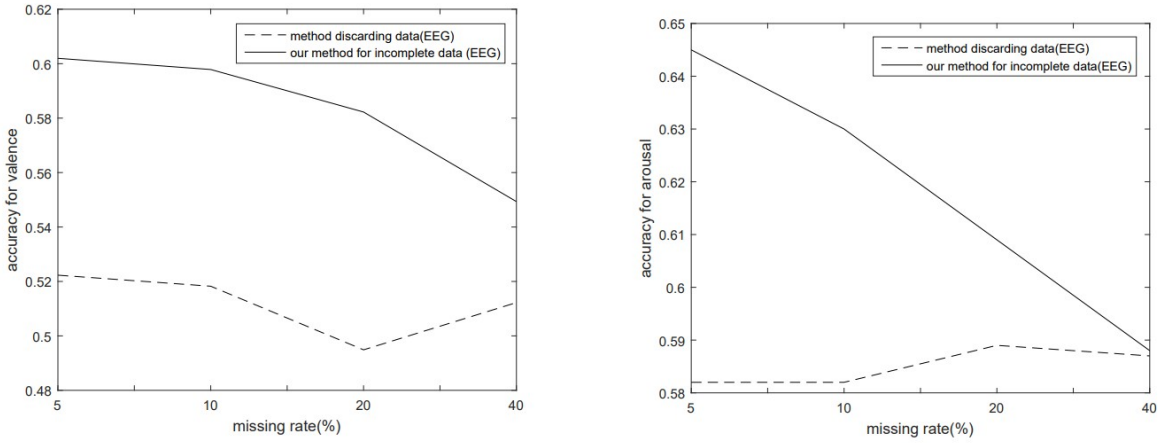

Figure 22. Graphs that represent comparison of Shu and Wagner's studies [1]

Second study [67], which was conducted after Wagner, involved use of EM algorithm, which is also maximum likelihood algorithm. Afterwards, restored data set was used in Bayesian network to make a decision.

7. EXPERIMENT

Based on the research, software that copes with missing data problem is created. Unfortunately, not all methods discussed in this thesis are implemented due to strict time limit on research. From all the methods, easiest ones to implement and test are chosen. Methods that are successfully implemented are:

1. Listwise deletion;
2. Mean substitution;
3. Cold deck imputation;
4. Hot deck imputation;
5. Linear regression analysis.

The software is created using C# language due to its simplicity and flexibility. Necessary debugging and performance measurement is conducted using standard debugging tools provided by Visual Studio. An example of working software is shown in Figure 23.

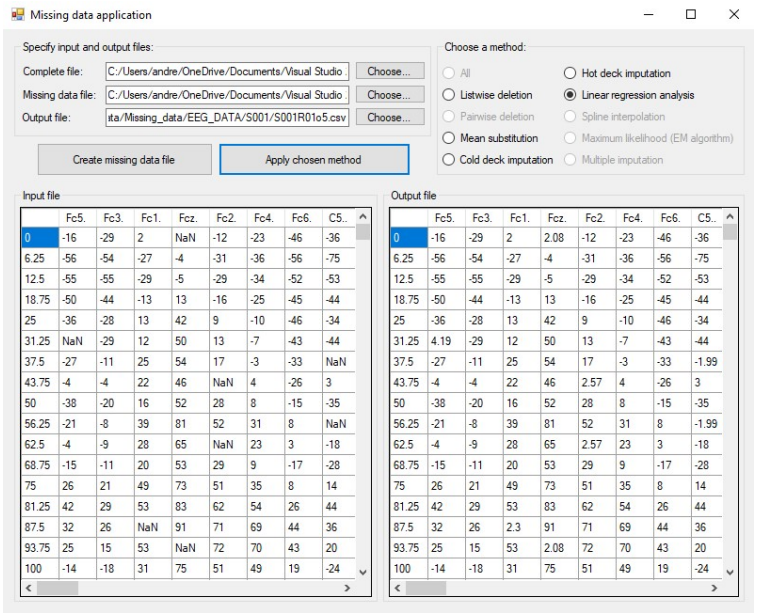

Figure 23. Software at work. In this example, linear regression analysis is used to restore the values.

Implemented functions for each method are available in Appendix. Full source code of an application is available online [69].

7.1. Data description

Data that is used to conduct the experiment is provided by [70-72]. The data itself is an EEG of 109 participants that were asked to do 14 different tasks. Each task took about 1-2 minutes to complete. Recordings of each of these tasks for each participant are located in a separate EDF file. Each of these files contains information about measurements of 64 electrodes located as shown in Figure 24.

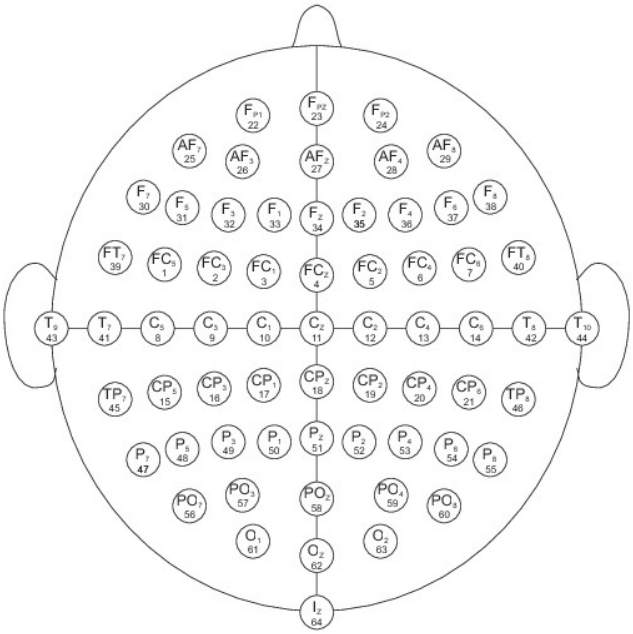

Figure 24. Location of electrodes [72]

Using EEGLAB software [73], these EDF files were converted to tab-delimited CSV files. This way, it is much easier to process them. Simplified structure of CSV files is shown in Table 6.

Table 6. Simplified structure of CSV files

Time (ms)	Electrode 1	Electrode 2	…	Electrode 64
0	-16	-29	…	25
6.25	-56	-54	…	36
…	…	…	…	…

These files contain complete data (no missing values are present). This is done on purpose to be able to compare this data to software's output to measure performance. Missing data files are created from these source files with 10% missingness (10% chance for each value to be lost). This way we can guarantee that we are dealing with MCAR data and all the methods are applicable.

All methods work only with missing data files and have no access to source files.

7.2. Results

After the data has been processed, results are analyzed. Listwise deletion had the worst results among all methods. As you can see from Figure 25, pretty much all the data is deleted while using this method. From original 9760 entries we only have 14 left. This happened due to relatively high missingness in files.

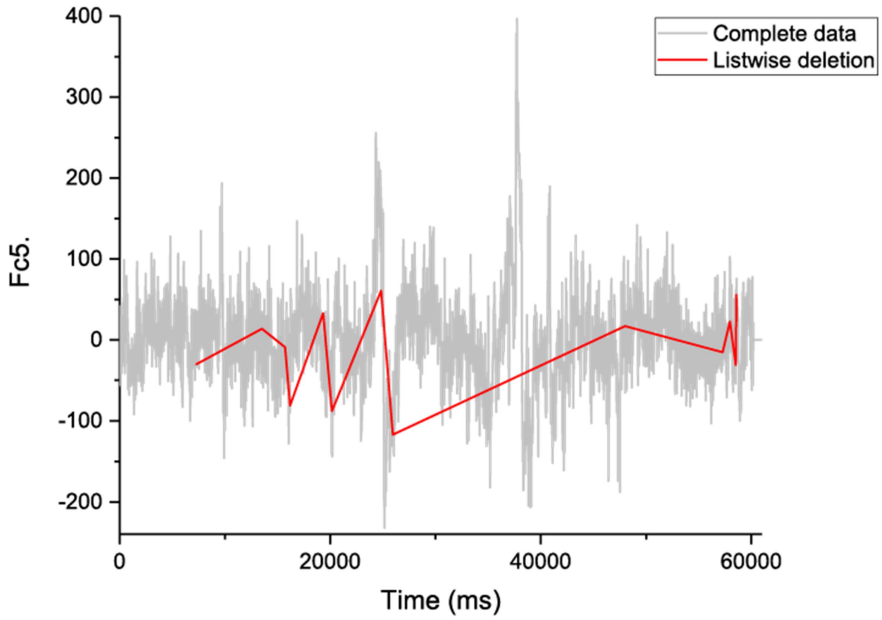

Figure 25. Listwise deletion analysis. Data is taken from electrode Fc5 for first task by first participant.

With 10% missingness every line only has 0.9^{64} chance to make it to output, which leaves us with 11.5 lines on average for every EDF file, which originally has 9760 lines. Even if we had only 1% missingness in files, almost 50% of data would be lost anyways. Obviously, these results cannot be treated seriously and this method should never be used if there is any risk to get more than 1% of missing data.

Imputation methods are much more promising. For imputation methods, least squares method is used to determine the best method. Main idea of least squares method is that sum

$$S = \sum_{i=1}^{n} (y_m - y_0)^2$$

should be minimized. Here y_m is value that has been imputed, y_0 is value that was actually acquired during the experiment, n is amount of missing values. The results are shown in Figure 26.

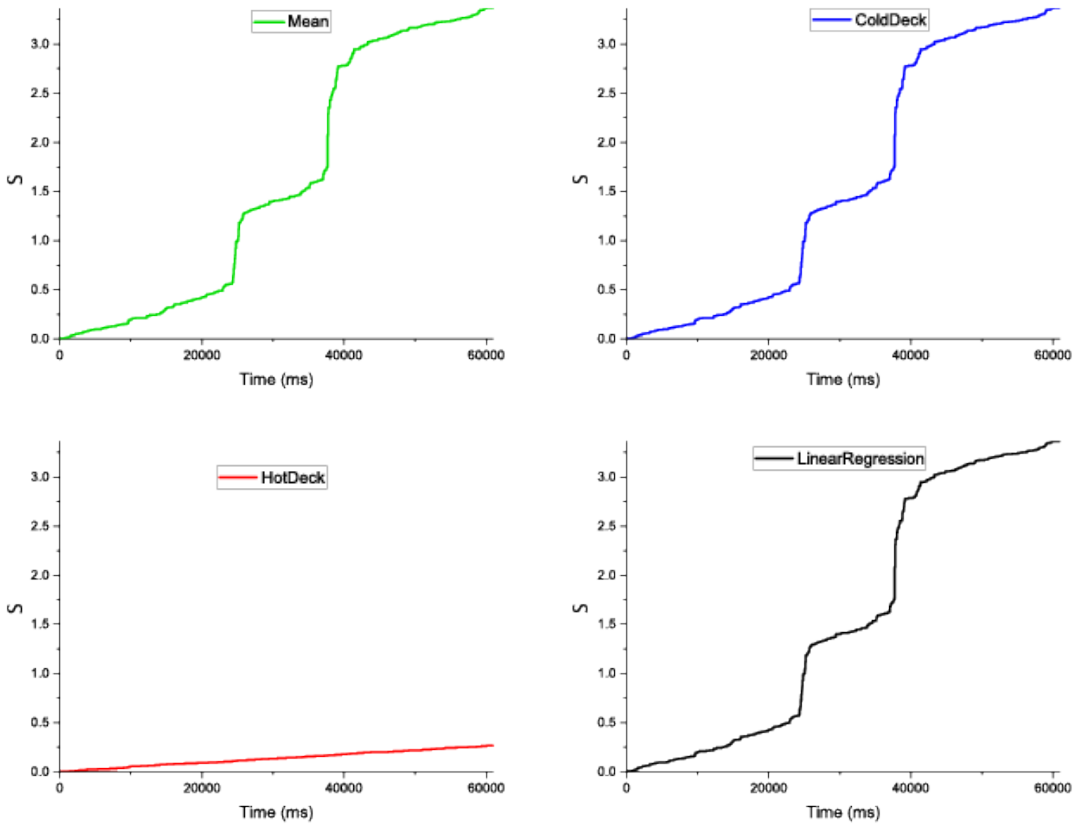

Figure 26. Analysis of imputation methods. Data is taken from electrode Fc5 for first task by first participant.

Interestingly enough, three methods showed very similar results. It can be explained that mean substitution, cold deck imputation and linear regression analysis are essentially doing the same thing. Due to the nature of EEG, all the values are centered around 0, which means that:

1. Mean value was 0;
2. Cold deck imputed 0;

3. Linear regression function is very similar to X axis.

This way, all of these three methods are imputing close to 0 values all the time. On the other hand, hot deck imputation method works differently; this is why his graph looks different comparing to others.

7.3. Performance measurements

During algorithm testing, performance of all implemented algorithms is also measured using debugging tools provided by Visual Studio. The conducted results are shown in Table 7.

Table 7. Performance measurements

	Listwise deletion	Mean substitution	Cold deck imputation	Hot deck imputation	Linear regression
Time (ms)	70	24	10	144048	33

As we can see, 4 of these methods are quick (less than a second), while hot deck imputation is rather slow (around 2 minutes).

8. CONCLUSION

The world of technology is rapidly developing. Only 50-60 years ago computers were size of a barn and their maintenance was quite expensive. Nowadays, we have insane (compare to what we had back then) computing power that can be and is used to make life of a researcher easier.

As it is already mentioned in introduction, missing data is a widespread fundamental problem that cannot be ignored. It distorts the data, sometimes even to the point where it is impossible to analyze data at all. That being said, we can use computers to take care of that problem for us (at least partially). However, some algorithms do require serious computing power even by today's standards; but nevertheless, it is still possible to use them in near future.

In emotion recognition, it is discovered that one of the best approaches to identify human emotions is by analyzing EEG results combined with peripheral signals. As you may know, EEG generates quite a lot of data and is subject to a problem of missing data. This way, it is extremely important to find out the methods to cope with missing data that would be applicable and efficient.

As this thesis has shown, there is no clear answer to this. You can say one thing for sure: listwise deletion method, which is so widely used and still is go-to strategy for most researchers, should never be used. Not only it removes huge chunks of data once missingness goes up making it impossible for data to be analyzed anyhow, it is not even cost-effective. It takes much less time to just use other methods instead, so what is even the point of it?

Three other methods (mean substitution, cold deck imputation and linear regression analysis) have showed almost identical results. It is, however, related specifically to the way EEG works. All EEG data is spread around 0 in such a way that mean value becomes 0 and linear regression function is very similar to X axis. This way, all missing values would be replaced with 0 or with very close to 0 numbers for all three cases. This brings us to the point that specifically for EEG it is actually a good idea to use cold deck imputation, because of its low demand on computer resources and decent results.

And finally, hot deck imputation. This method proved to be the best of all tested methods. However, it is also the most demanding of them. This means that if you have an access to good computer, or your amount of data is relatively short so speed would not be an issue, then this method is the best choice.

8.1. Answers to research questions

During the study, all 3 research questions have been answered. Let us shortly summarize all the answers.

RQ1. What are the techniques that can be used to deal with missing data?

The techniques that can be used for this purpose are: listwise deletion, pairwise deletion, mean substitution, cold deck imputation, hot deck imputation, linear regression analysis, spline interpolation, maximum likelihood estimation and multiple imputation.

RQ2. Which techniques are the most suitable to use in emotion recognition?

Out of all tested techniques (see Chapter 7), most suitable techniques for emotion recognition are mean substitution, cold deck imputation, hot deck imputation and linear regression analysis.

RQ3. When each of those suitable techniques can be used?

If you have enough computer power or small amount of data, then you should definitely use hot deck imputation method. Otherwise, mean substitution, cold deck imputation and linear regression analysis are viable methods to use in emotion recognition.

8.2. Future works

It is worth noting that not all available methods are tested during this research. Pairwise deletion, spline interpolation, EM algorithm and multiple imputation methods should all be

tested. It might be one of them that is more suitable, than what we have now. Also, some methods that are not included in thesis (they are discussed in Section 3.2) should be tested as well.

Moreover, these methods should be tested under big pressure as well. Using large amounts of data to process might also change behavior of these methods. So it is definitely something that should be done in future.

Considering the level of technology that we have now, there is no real reason not to handle missing data. At very least, every researcher who is working with any kind of data acquisition, should know the basic ways to handle it. It will allow reducing amount of mistakes that are made because of incorrect data analysis and thus make the researches more consistent and reliable. Works similar to this one should be done in every field of science to determine the best suitable techniques for researchers to use. This way, we can build a solid basis which will help future scientists to conduct their researches more reliably.

REFERENCES

[1] Y. Shu and S. Wang, "Emotion recognition through integrating EEG and peripheral signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 2871-2875.

[2] J. L. Lakin and T. L. Chartrand, "Using nonconscious behavioral mimicry to create affiliation and rapport" *Psychological science,* vol. 14, pp. 334-339, 2003.

[3] C. C. Chibelushi and F. Bourel, "Facial expression recognition: A brief tutorial overview" *CVonline: On-Line Compendium of Computer Vision,* vol. 9, 2003.

[4] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems" *ACM Computing Surveys (CSUR),* vol. 47, p. 43, 2015.

[5] Mayo Clinic. (2018). *EEG*. Available: https://www.mayoclinic.org/tests-procedures/eeg/about/pac-20393875

[6] Merriam-Webster. (2018). *Magnetoencephalography*. Available: www.merriam-webster.com/dictionary/magnetoencephalography

[7] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain" *Reviews of modern Physics,* vol. 65, p. 413, 1993.

[8] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*: Lippincott Williams & Wilkins, 2005.

[9] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch*, et al.*, "Towards personal service robots for the elderly," in *Workshop on Interactive Robots and Entertainment (WIRE 2000)*, 2000, p. 184.

[10] M. Farsi, M. Munro, and A. Al-Thobaiti, "The effects of teaching primary school children the Islamic prayer in a virtual environment," in *Science and Information Conference (SAI), 2015*, 2015, pp. 765-769.

[11] OECD. (2005, 19th of December). *OECD Glossary of Statistical Terms - Missing data definition*. Available: https://stats.oecd.org/glossary/detail.asp?ID=6131

[12] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data" *IEEE Transactions on Affective Computing,* vol. 2, pp. 206-218, 2011.

[13] J. Wagner, F. Lingenfelser, and E. André, "Building a robust system for multimodal emotion recognition" *Emotion recognition: A pattern analysis approach,* pp. 379-410, 2015.

[14] F. R. Al-Osaimi, M. Bennamoun, and A. Mian, "Integration of local and global geometrical cues for 3D face recognition" *Pattern Recognition,* vol. 41, pp. 1030-1040, 2008.

[15] L. Ballihi, B. B. Amor, M. Daoudi, A. Srivastava, and D. Aboutajdine, "Boosting 3-D-geometric features for efficient face recognition and gender classification" *IEEE Transactions on Information Forensics and Security,* vol. 7, pp. 1766-1779, 2012.

[16] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "A region ensemble for 3-D face recognition" *IEEE Transactions on Information Forensics and Security,* vol. 3, pp. 62-73, 2008.

[17] S. Feng, H. Krim, and I. Kogan, "3D face recognition using Euclidean integral invariants signature," in *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, 2007, pp. 156-160.
[18] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, "3D face recognition using covariance based descriptors" *Pattern Recognition Letters,* vol. 78, pp. 1-7, 2016.
[19] S. Jahanbin, H. Choi, Y. Liu, and A. C. Bovik, "Three dimensional face recognition using iso-geodesic and iso-depth curves," in *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, 2008, pp. 1-6.
[20] Y. Lei, M. Bennamoun, and A. A. El-Sallam, "An efficient 3D face recognition approach based on the fusion of novel local low-level features" *Pattern Recognition,* vol. 46, pp. 24-37, 2013.
[21] Y. Lei, M. Bennamoun, M. Hayat, and Y. Guo, "An efficient 3D face recognition approach using local geometrical signatures" *Pattern Recognition,* vol. 47, pp. 509-524, 2014.
[22] H. Li, D. Huang, J.-M. Morvan, L. Chen, and Y. Wang, "Expression-robust 3D face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns" *Neurocomputing,* vol. 133, pp. 179-193, 2014.
[23] X. Li and H. Zhang, "Adapting geometric attributes for expression-invariant 3D face recognition," in *Shape Modeling and Applications, 2007. SMI'07. IEEE International Conference on*, 2007, pp. 21-32.
[24] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen, "An intrinsic framework for analysis of facial surfaces" *International Journal of Computer Vision,* vol. 82, pp. 80-95, 2009.
[25] S. Soltanpour, B. Boufama, and Q. J. Wu, "A survey of local feature methods for 3D face recognition" *Pattern Recognition,* vol. 72, pp. 391-406, 2017.
[26] H. Tabia, H. Laga, D. Picard, and P.-H. Gosselin, "Covariance descriptors for 3D shape matching and retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4185-4192.
[27] A. Knutas, A. Hajikhani, J. Salminen, J. Ikonen, and J. Porras, "Cloud-based bibliometric analysis service for systematic mapping studies," in *Proceedings of the 16th International Conference on Computer Systems and Technologies*, 2015, pp. 184-191.
[28] R. J. Little and D. B. Rubin, "The analysis of social science data with missing values" *Sociological Methods & Research,* vol. 18, pp. 292-326, 1989.
[29] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing incomplete political science data: An alternative algorithm for multiple imputation" *American political science review,* vol. 95, pp. 49-69, 2001.
[30] P. L. Roth, "Missing data: A conceptual review for applied psychologists" *Personnel psychology,* vol. 47, pp. 537-560, 1994.
[31] J. L. Peugh and C. K. Enders, "Missing data in educational research: A review of reporting practices and suggestions for improvement" *Review of educational research,* vol. 74, pp. 525-556, 2004.

[32] O. Harel, R. Zimmerman, and O. Dekhtyar, *Approaches to the handling of missing data in communication research*: University of Connecticut, Department of Statistics, 2007.
[33] Pogodaiklimat. (2018). *Climate in Saint Petersburg (In Russian)*. Available: http://www.pogodaiklimat.ru/climate/26063.htm
[34] D. B. Rubin, "Inference and missing data" *Biometrika,* vol. 63, pp. 581-592, 1976.
[35] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987.
[36] B. Muthén, D. Kaplan, and M. Hollis, "On structural equation modeling with data that are not missing completely at random" *Psychometrika,* vol. 52, pp. 431-462, 1987.
[37] T. E. Raghunathan, "What do we do with missing data? Some options for analysis of incomplete data" *Annu. Rev. Public Health,* vol. 25, pp. 99-117, 2004.
[38] P. D. Allison, "Missing data: Quantitative applications in the social sciences" *British Journal of Mathematical and Statistical Psychology,* vol. 55, pp. 193-196, 2002.
[39] C. K. Enders, *Applied missing data analysis*: Guilford Press, 2010.
[40] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art" *Psychological methods,* vol. 7, p. 147, 2002.
[41] J. W. Graham, "Missing data analysis: Making it work in the real world" *Annual review of psychology,* vol. 60, pp. 549-576, 2009.
[42] R. J. Little, "A test of missing completely at random for multivariate data with missing values" *Journal of the American Statistical Association,* vol. 83, pp. 1198-1202, 1988.
[43] N. Zagoruiko. *ZET algorithm for filling gaps*. Available: http://math.nsc.ru/AP/oteks/English/begin/links/Zet/index.html
[44] N. Zagoruiko, V. Elkina, and V. Temirkaev, "ZET—An algorithm of filling gaps in experimental data tables" (In Russian), *Comput. Syst,* vol. 67, pp. 3-28, 1976.
[45] N. Zagoruyko, V. Elkina, and V. Timerkaev, "Algorithm for filling gaps in empirical tables (algorithm Zet)" (In Russian), *Vychislitel'nye sistemy: sb. tr. Vyp. 61. Empiricheskoe predskazanie i raspoznavanie obrazov,* pp. 3-27, 1975.
[46] J. W. Graham and S. I. Donaldson, "Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data" *Journal of Applied Psychology,* vol. 78, p. 119, 1993.
[47] W. Wothke, "Longitudinal and multigroup modeling with missing data" 2000.
[48] T. D. Pigott, "A review of methods for missing data" *Educational research and evaluation,* vol. 7, pp. 353-383, 2001.
[49] T. A. Myers, "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data" *Communication Methods and Measures,* vol. 5, pp. 297-310, 2011.
[50] A. C. Acock, "Working with missing values" *Journal of Marriage and family,* vol. 67, pp. 1012-1028, 2005.
[51] R. J. Little, "Regression with missing X's: a review" *Journal of the American Statistical Association,* vol. 87, pp. 1227-1237, 1992.
[52] D. A. Bennett, "How can I deal with missing data in my study?" *Australian and New Zealand journal of public health,* vol. 25, pp. 464-469, 2001.

[53] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response" *International statistical review,* vol. 78, pp. 40-64, 2010.
[54] N. R. Draper and H. Smith, *Applied regression analysis* vol. 326: John Wiley & Sons, 2014.
[55] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis* vol. 821: John Wiley & Sons, 2012.
[56] J. W. Gorman and R. Toman, "Selection of variables for fitting equations to data" *Technometrics,* vol. 8, pp. 27-51, 1966.
[57] M. Box and N. R. Draper, "Factorial designs, the| X′ X| criterion, and some related matters" *Technometrics,* vol. 13, pp. 731-742, 1971.
[58] S. Wold, A. Ruhe, H. Wold, and I. Dunn, WJ, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses" *SIAM Journal on Scientific and Statistical Computing,* vol. 5, pp. 735-743, 1984.
[59] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor, *A practical guide to splines* vol. 27: Springer-Verlag New York, 1978.
[60] D. F. Rogers and J. A. Adams, *Mathematical elements for computer graphics*: McGraw-Hill Higher Education, 1989.
[61] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm" *Journal of the royal statistical society. Series B (methodological),* pp. 1-38, 1977.
[62] C. J. Wu, "On the convergence properties of the EM algorithm" *The Annals of statistics,* pp. 95-103, 1983.
[63] A. Riley. (2017). *EM algorithm*. Available: https://stackoverflow.com/questions/11808074/what-is-an-intuitive-explanation-of-the-expectation-maximization-technique
[64] J. L. Schafer, *Analysis of incomplete multivariate data*: CRC press, 1997.
[65] J. L. Schafer, "Multiple imputation: a primer" *Statistical methods in medical research,* vol. 8, pp. 3-15, 1999.
[66] J. A. Russell, "A circumplex model of affect" *Journal of personality and social psychology,* vol. 39, p. 1161, 1980.
[67] I. Cohen, N. Sebe, F. Gozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2003, pp. I-I.
[68] D. Graziotin, X. Wang, and P. Abrahamsson, "Understanding the affect of developers: theoretical background and guidelines for psychoempirical software engineering," in *Proceedings of the 7th International Workshop on Social Software Engineering*, 2015, pp. 25-32.
[69] A. Gorbulin. (2018). *Missing data*. Available: https://github.com/host-ru/Missing-data
[70] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, *et al.*, "Physiobank, physiotoolkit, and physionet" *Circulation,* vol. 101, pp. e215-e220, 2000.

[71] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system" *IEEE Transactions on biomedical engineering,* vol. 51, pp. 1034-1043, 2004.
[72] Physionet. (2009). *EEG Motor Movement/Imagery Dataset*. Available: https://www.physionet.org/pn4/eegmmidb/
[73] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis" *Journal of neuroscience methods,* vol. 134, pp. 9-21, 2004.

APPENDIX

```
public void ListwiseDeletion()
{
    outputList = inputList;

    for (int i = 0; i < outputList.Count; i++)
    {
        for (int j = 0; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                outputList.RemoveAt(i);
                i--;
                break;
            }
        }
    }
}

public void MeanSubstitution()
{
    outputList = inputList;

    double[] mean = new double[outputList[0].Length];
    double[] n = new double[outputList[0].Length];

    // Calculate mean values
    for (int i = 0; i < outputList.Count; i++)
    {
        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                continue;
            }
            else
            {
                mean[j] += outputList[i][j];
                n[j]++;
            }
        }
    }

    for (int i = 1; i < mean.Length; i++)
    {
        mean[i] /= n[i];
        mean[i] = Math.Round(mean[i], 2);
    }

    // Impute mean values istead of NaN
    for (int i = 0; i < outputList.Count; i++)
    {
```

```
        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                outputList[i][j] = mean[j];
            }
        }
    }
}

public void ColdDeckImputation()
{
    outputList = inputList;

    for (int i = 0; i < outputList.Count; i++)
    {
        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                outputList[i][j] = 0;
            }
        }
    }
}

public void HotDeckImputation()
{
    outputList = inputList;

    for (int i = 0; i < outputList.Count; i++)
    {
        double[] squareSum = new double[outputList.Count];

        // Calculate squareSum for each row in relation to this row
        for (int k = 0; k < outputList.Count; k++)
        {
            for (int l = 1; l < outputList[i].Length; l++)
            {
                if (double.IsNaN(outputList[k][l]))
                {
                    squareSum[k] = double.MaxValue;
                    break;
                }
                else if (double.IsNaN(outputList[i][l]))
                {
                    continue;
                }
                else
                {
                    squareSum[k] += (outputList[k][l] - outputList[i][l]) *
(outputList[k][l] - outputList[i][l]);
                }
            }
```

```
        }

        // Use the one with least square sum
        int leastSquareSumIndex = Array.IndexOf(squareSum, squareSum.Min());

        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                outputList[i][j] = outputList[leastSquareSumIndex][j];
            }
        }
    }
}

public void LinearRegression()
{
    outputList = inputList;

    // Calculate necessary parameters for each electrode
    double[] a = new double[outputList[0].Length];
    double[] b = new double[outputList[0].Length];
    double sumX = 0;
    double[] sumY = new double[outputList[0].Length];
    double[] sumXY = new double[outputList[0].Length];
    double sumX2 = 0;
    double[] n = new double[outputList[0].Length];

    for (int i = 0; i < outputList.Count; i++)
    {
        sumX += outputList[i][0];
        sumX2 += outputList[i][0] * outputList[i][0];

        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                continue;
            }
            else
            {
                sumY[j] += outputList[i][j];
                sumXY[j] += outputList[i][0] * outputList[i][j];
                n[j]++;
            }
        }
    }

    // Calculate coefficients a and b
    for (int j = 1; j < outputList[0].Length; j++)
    {
        b[j] = (n[j] * sumXY[j] - sumX * sumY[j]) / (n[j] * sumX2 - sumX * sumX);
        a[j] = (sumY[j] - b[j] * sumX) / n[j];
    }
```

```
    // Use y = a + bx to substitute missing values
    for (int i = 0; i < outputList.Count; i++)
    {
        for (int j = 1; j < outputList[i].Length; j++)
        {
            if (double.IsNaN(outputList[i][j]))
            {
                outputList[i][j] = Math.Round(a[j] + b[j] * outputList[i][0], 2);
            }
        }
    }
}
```