Sustainable Income Algorithm: Finpass

Lina Dindiene[1], Audrius Kabasinskas[1], Armin Kruppy[2], Victoria Pereira[2]†, and Bogdan Toader[2].

[1] *Kaunas University of Technology, Lithuania*
[2] *Mathematical Institute, University of Oxford, UK*

(*Communicated to* MIIR *on 5 November 2021*)

Study Group: ESGI142, June 11–15, 2018, Palanga, Lithuania

Communicated by: Chris Breward

Industrial Partner: Finpass

Team Members: Lina Dindiene, Vilnius University; Audrius Kabasinskas, Kaunas University of Technology; Armin Kruppy, University of Oxford; Victoria Pereira, University of Oxford; Bogdan Toader, University of Oxford.

Industrial Sector: Finance

Key Words: Income Prediction, Responsible lending, Consumer credit, Systemic risk,

MSC2020 Codes: 62

Summary

Finpass is a finance technology start-up. In this report, we present exploratory data analysis of historical bank transactions, and developed an algorithm to predict future sustainable income.

† Corresponding Author: `Victoria.Pereira@maths.ox.ac.uk`

Sustainable Income Algorithm: Finpass

Lina Dindiene*, Audrius Kabasinskas*, Armin Krupp†,
Victoria Pereira†, Bogdan Toader†

June 2018

Abstract

In Lithuania, financial institutions that lend money are regulated to include *sustainable income* in their calculations for the amount that they are able to lend. Typically most calculations of sustainable income are based on the mean of a salary or pension payments. There are however different types of regular income sources including bills and rental income. In this Study Group Report we use anonymous bank transaction data to develop a data-driven algorithm for predicting an individuals' sustainable outcome.

1 Introduction and problem overview

In 2011 the Bank of Lithuania first introduced the *Responsible Lending Regulations* which sought to encourage the practice of responsible lending by credit institutions, maintain the market's discipline and ensure transparency of operations, and decrease the systemic risk of the credit institutions sector thus helping to ensure the stability of the financial system. The Responsible Lending Regulations obligate credit institutions to fully assess the ability of credit borrowers to return credit in the long term, define the largest permitted loan-to-value ratio as well as the largest debt-service-to-income (DSTI) ratio. The regulations state that *sustainable income* must be included in the calculate of the DSTI ratio. However, the Bank of Lithuania does not provide an exact definition of sustainable income and leaves it to lenders to decide what income can be recognised as sustainable.

In general, sustainable income is understood to be the income that a lender can expect a borrower to receive over a period of a loan. The Bank of Lithuania talks about sustainable income in the following regulatory papers:

1. Responsible lending regulations (RLR)
2. Responsible lending regulations for consumer credit (RLR for CC)
3. Guidelines on issuing consumer credit (guidelines)

The RLR and RLR for CC have a minor, yet very important difference. RLR defines sustainable income for the particular case of a borrower applying for a mortgage. In this instance, the RLR states that income from at least 6 previous months should be taken in account when determining sustainable income. The RLR for CC defines sustainable income for the setup where a borrower applies for a consumer loan. In this case, income from at least 4 previous months should be taken in account when determining sustainable income. This difference in time period can have significant implications in the calculation of sustainable income. For example consider a person that only has employment history of 5 months with his first employee. His income would be considered sustainable for consumer loan, but not for mortgage.

1.1 Current practice

There is no set definition of sustainable income. The Bank of Lithuania does propose lenders use an 'average-based model', but the current practice is for each lender to use its own proprietary methodology to calculate sustainable income. This is a time-consuming and labour-intensive task, that should be standardised and automated as much as possible.

Moreover, as there is no uniform methodology on calculating sustainable income, lenders have a considerable amount of leeway with regards to what type of income should be treated as sustainable income. However, it is also up to lenders to prove that their calculations conform to regulatory requirements if they are inspected by the Bank of Lithuania.

[1] Kaunas University of Technology, Lithuania
[2] University of Oxford, UK

The only automated process that can be used to check for sustainable income is social security (i.e. SoDra) registry, which holds information on income that is used for social security taxation purposes. However, this registry does not keep income records for those individuals who are self employed, or receive income from other sources, such as rent or dividends (i.e. income not used is social security taxation calculation).

According to the information regarding the composition of the monthly disposable income in cash provided by the Lithuanian Department of Statistics, the income from employment salaries makes up about 60 percent of all disposable income (per household member in 2014, 2015 and 2016 years). More precisely, in cities this number is approximately 65 %, while in the countryside it is about 50 %. Therefore, a significant part of the income comes from alternative sources which banks currently do not include in calculations of sustainable income. A consequent of this is that banks are overly strict for individuals who do not have regular salary payments, but do have alternative stable sources of income.

Some lenders also use raw or categorised client's bank statement data in order to determine what types of income their client earns in addition to those that are used for social security calculations. Yet, this source provides only additional information, but no actual solution.

A number or lenders use and pay for both sources of income information: social security registry and bank statement data. However, in most cases social security income information is already present in bank statement data. Thus lenders are paying twice for essentially the same information and receive no complete answer.

1.2 Project outline

In this Study Group, our objective is to create a sustainable income calculation algorithm that complies with the Bank of Lithuania RLRs & Guidelines and uses information only from a categorised bank statement (i.e. categorised transactions). This project was brought to the Study Group by the Fintech company Finpass. They seek an algorithm that is 'forward looking', that is it should not be limited to calculating average historic sustainable income and it should be able to forecast sustainable income for the near future, i.e. duration of a consumer loan.

The algorithm should have the following calculation options:

1. Sustainable income calculation for mortgages (minimum of 6 month income information used)
2. Sustainable income calculation for consumer credit (minimum of 4 month income information used
3. Sustainable income calculations where some types of incomes are excluded
4. Sustainable income calculations based on all types of income

Moreover, Finpass advised that income sources that must be included in the algorithm are:

1. Pension
2. Salary

This following transactions of regular or recurring bills should also be included:

1. Bills
2. Dividends received
3. Interest received

The regularity or sustainable of income from alternative sources needs to be determined. This sources include

1. Alimony and child support
2. Online payments
3. Other incoming payments from employer (OIP from employer)
4. Rental income
5. Royalty fee
6. Scholarships
7. Securities
8. Advance payments
9. Social security benefits

10. Tax return

Finally, income from the following sources should not be treated as sustainable:

1. Advance payment
2. Cash deposit
3. Currency exchange
4. Deposits
5. Gambling
6. Insurance indemnity
7. Loans (student loans, consumer loans, leasing, etc.)
8. Online money transfer
9. Other incoming payments
10. Personal transfers (between accounts, from relatives, other)
11. Returned payments
12. Savings
13. State aid payments

In this report, we first present an exploratory data analysis of the dataset in which we analyse each type of income separately. We consider clusters of individuals, and derive forecasting models to predict stable income based on a number of approaches including a weighted median based model and an autoregressive model. We assess the performance of our forecast by comparing prediction to real life data.

2 Exploratory Data Analysis

The goal of the exploratory data analysis is to understand the dataset, motivate the creation and selection of features, and provide an idea of the kind of models that could work.

This section consists of four subsections. In Section 2.1 we will first analyse the full dataset and consider the distribution of transaction frequency over time. This analysis will be useful to guide the model testing and selection. In the Section 2.2, we will look at the distribution of monthly incomes from sustainable sources and how they vary, measured by the standard deviation. In Section 2.3, we will analyse the importance of different income streams for different people. This will allow us to understand the impact more complex models than a simple income average will have. In Section 2.4, we consider clustering of the account holders to provide a reduced parameter space for prediction, that is can we predict what groups of individuals rather than predicting on the individual level.

2.1 Exploring the dataset: Volume, Frequency, and active users

For each individual, their bank statements constitutes a time series, which can be used to predict the income from different income sources. While certain income sources, such as salary, pensions, or benefits occur with a certain regularity such as the last day of the month, the fact that the months have different lengths makes it impractical to consider, say, the day of the month as the time component of the time series. Instead, we aggregate all transactions from valid income sources per month, and will hence use the month of the year as the time unit for our time series.

In figure 1, we plot the number of all and valid transactions per month for our dataset. In the dataset, around a third of all transactions are from valid sources, meaning that there are typically between $5,000$ to $10,000$ transactions to consider in our model. For typical lookback periods of a year, this corresponds to approximately $100,000$ datapoints.

While there is limited data available pre-2008, we see a steady increase in the number of transactions leading up to 2015 and a significant jump in activity until the beginning of 2018. This suggests that the dataset is not complete from 2018 onwards. Thus we will work only with the time period of 1/1/2015 – 31/12/2017 in this report.

In figure 2, we display the average number of valid transactions per account holder. We use the first time a given account holder received any income to their account to define the first month the account was active

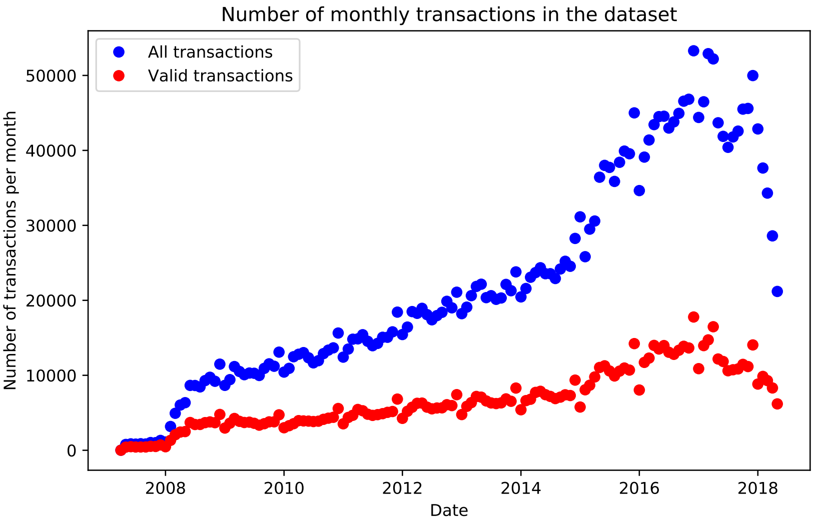


Figure 1: Number of incoming transactions from all (blue) and valid (red) sources in the dataset.

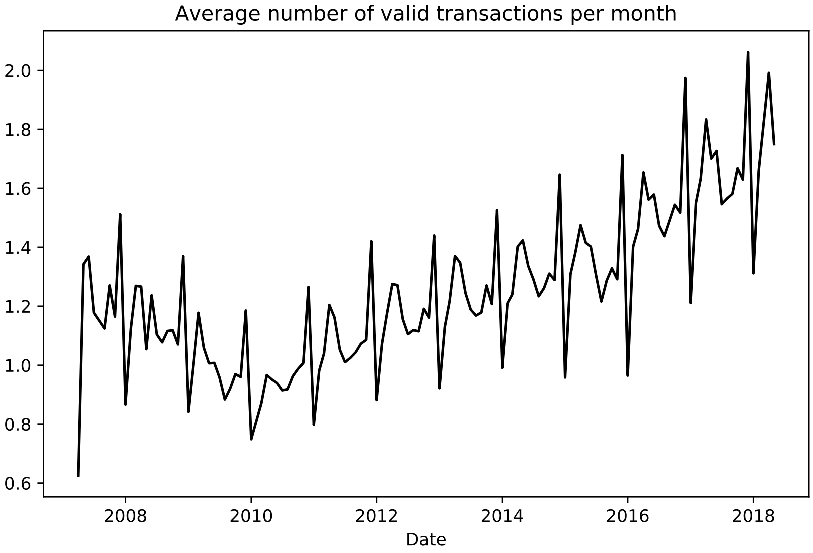


Figure 2: Average number of valid transactions per account holder

and equally the last time an account received any income (not necessarily from valid income sources) to define the last month the account was active. We then use a multiindex to define all active months so that months in between the first and last active months where no transactions happened are included as zeros.

On this basis, we observe that the mean number of transactions is very low, hence for the most account holders, only 1 or 2 data points will be available per month. We summarise the median quantiles for the years 2016 and 2017 in Table 1.

Table 1: Median quantile values for the number of valid income transactions per month for 2016 and 2017.

Quantile	25%	Median	75%	90%	95%	99%
Value	1	1	2	3	4	6

As the mean of the number of transactions suggests, the typical number of valid income transactions is only around 1 (the median) and only is larger or equal to 4 for the top 5%. Thus we have a low number of valid transactions for each individual.

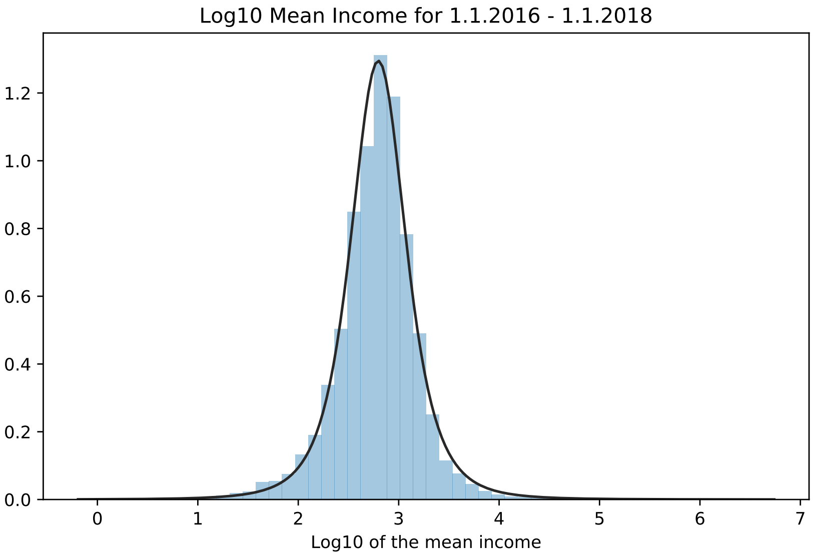


Figure 3: Distribution of the $\log_{10}(x + 1)$ of the monthly incomes from valid sources. The distribution of the data is very well approximated by a normal distribution

2.2 Sustainable income distribution in the dataset: How are they distributed?

Looking at the distribution of monthly income, our exploratory data analysis suggests that the distribution of the median monthly income as well as the the standard distribution are both approximately constant over time. We show the distribution (base 10 logarithm) of the monthly incomes from valid sources in figure 3. We show a distribution plot for the base 10 logarithm and add 1 in order to cope with zeros in the dataset. From this we see that the logarithm of the income is very well approximated by a normal distribution, exhibiting a slight skew to the right. The mean of the corresponding normal distribution is 2.78, corresponding to an income (including the +1 of 602.79 euros per month, the standard deviation of the logarithms is $\sigma = 0.4219$).

2.3 The relative importance of different income streams

The data provides us with 15 different income types that can count towards computing the sustainable income (see Section 1.2 for details). To understand how many different income streams will be considered per account holder, we determine the distribution of the number of different income streams the account holders have in 2017, the results are shown in figure 4.

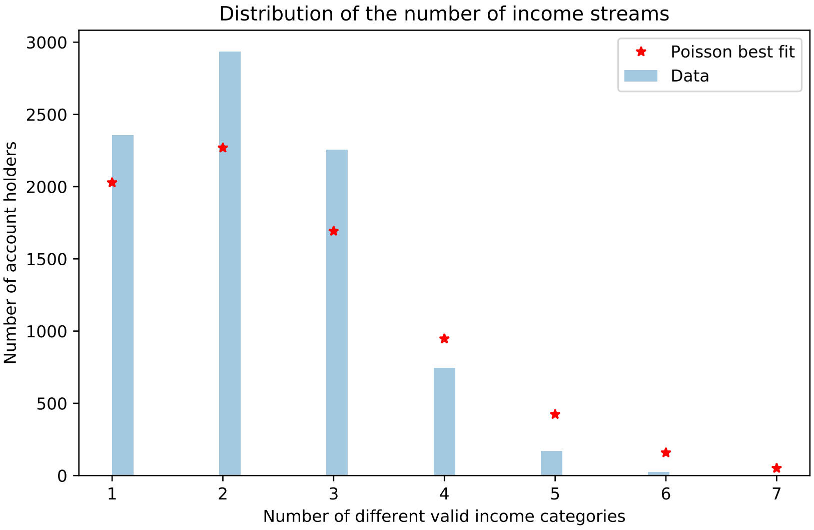


Figure 4: Distribution of the number of different income streams per account holder in 2017, the distribution is reasonably approximated by the Poisson-distribution.

In line with the results from 2, we see that the majority of the account holders only have up to 3 different income streams over the entire year, whereas a minority has 4 or more. We also find that the Poisson-distribution reasonably approximates the distribution of the number of different income streams.

To better understand the relevance of the different income streams, we scatter the average monthly income for a given stream (with the average taken only by account holders who receive a non-zero income from said income stream) against the number of people receive money via said stream, the results are shown in figure 5.

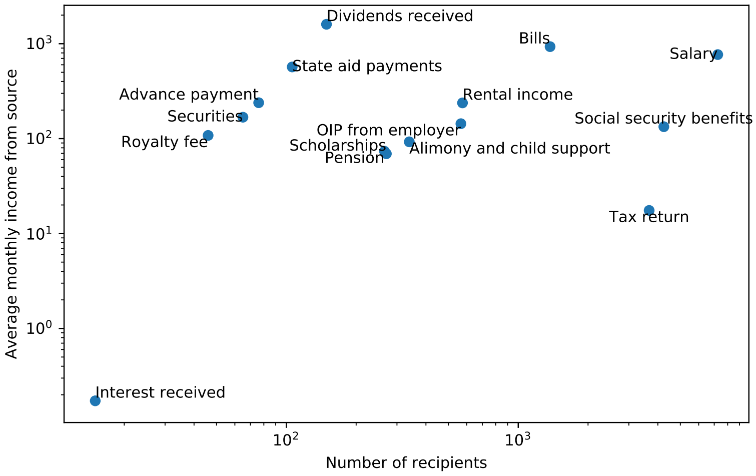


Figure 5: Scatter plot for the different income streams based on their monthly average volume (only taking into account non-zero contributors) and the number of people that receive income from said stream in 2017.

Four different clusters emerge, with three income sources (*salary, social security benefits*, and *tax return*) being relevant to more than 1000 people, 5 to more than hundred, and the rest only being relevant to a small group of people. *Receiving interest* has the additional peculiarity that it is comparatively small in size.

2.4 Clustering of individuals

Where we have a high number of individuals and transaction we can consider the use of clustering to reduce the dimensionality of the problem. Clustering is used to partition the data into groups, where data points in the groups are closer to each other than they are to other groups. In this data set, we can consider clustering in many ways:

1. Cluster on the client level over all their transactions across different categories.
2. Cluster on the client level over their transaction within a single category.
3. Cluster over all the clients based on transactions within a single category.

We ran k-means clustering for these three clustering objectives, but found little cluster structure. For example, see figure 6. This shows the similarity matrix ordered by clusters for client 422. In this example we are seeking cluster of similar transactions for the single client. If there was successful clustering, clear block matrices would be seen. The results suggest that clustering the data in these ways is not helpful to reduce the size of the problem.

2.5 Summary and implications

In this section we have shown that while there are 15 different income streams, for a very large majority only a few, namely their salary and/or social security benefits will be relevant. This implies that the baseline model of only including the salary into the calculation of the sustainable income will be sufficient for most people, and we should only expect to improve the baseline model in 10% or so of the cases. Additionally, we must be careful about how to sample and validate when building models for the less frequent income streams.

The fact that the quality of the data is good for the period of 2015 until the end of 2018 will be used to train models on 2015 for predicting 2016 and using 2016 to predict 2017 as the validation.

3 Predictive models

In this section, we discuss and introduce a number of approaches to forecast sustainable income based on historic transactional data. In Section 3.1 we discuss the general form and motivation of a weighted mean/median model.

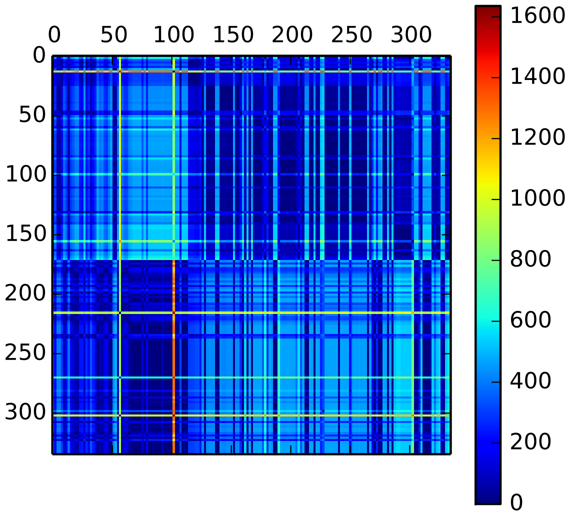


Figure 6: Clusters for client 422 over all categories.

In Section 3.2, we present a median-based model with quantile factors. In Section 3.3, we fit an autoregressive model to the income data to capture the periodicity present in the data.

3.1 Weighted mean/median based models

Different incomes are incoming with different periodicity. Some of them are regular and some are irregular. A weighted mean/median model has the general form:

$$Y_t = \sum_{i=1}^{n} a_i \cdot f^{(i)}(x_{t-1}^{(i)}, x_{t-2}^{(i)}, ...), \tag{1}$$

here a_i are regularity weights calculated as rate of months with occurrence of the income over total number of months taken into account; $x_{t-1}^{(i)}, x_{t-2}^{(i)}, ...$ are past incoming payments of type i; function $f^{(i)}(\cdot)$ aggregates past information about the income, could be monthly mean, median or other quantiles, exponentially smoothed mean etc.

Name of the model	applied for	calculation	notes
mean	all	mean of category for each month	
median	all	median of category for each month	
upper bound of CI	very regular income	mean/median+error	in general 1 std dev
quantile regression estimate	regular income	estimate from the model	in case of median regression it is the same as 2d model
time series	regular income	estimate from the model	
exponentially weighted mean	all	estimate from the model	e.g. last 3 months has higher importance

Table 2: Possible choices of aggregation function $f^{(i)}(\cdot)$

For example Statement Holder 422, male of age 56. Has three types of income (Cash deposit, Personal transfers, Salary) that appear 2, 6 and 12 in last 12 months. Their average and median values (excluding months when they haven't appeared) are given in table 3

type of income	occurrence	mean (€)	median(€)
Cash deposit	2	250	250
Personal transfers	6	137.889	45
Salary	12	1204.605	1079.93

Table 3: statement holder information

Then the formula for calculation of stable income from means is:

$$Y_t^{mean} = 2/12 \cdot 250 + 6/12 \cdot 137.889 + 12/12 \cdot 1204.605 = 1315.216€ \quad (2)$$

Then the formula for calculation of stable income from medians is:

$$Y_t^{median} = 2/12 \cdot 250 + 6/12 \cdot 45 + 12/12 \cdot 1079.93 = 1144.097€ \quad (3)$$

It must be noticed that median estimate is not so sensitive to extreme cases of incoming payments. However if we use only average incomes that may be treated as stable/sustainable in that case we get only 1238.6375 €.

3.1.1 Confidence intervals based models

The confidence interval for median is constructed in following way:

$$CI_{t,i} = median_{t,i} \pm \sigma_{t,i} \cdot k \cdot stdmed(\alpha, n)_i \quad (4)$$

here $stmded(\alpha, n)$ is standard error of median and is found by methodology described by SAS software [2] or a very robust estimate for observations $n > 5$ months.

$$stmded(n, \alpha) = n - (floor(n/2) - ceil(t_{\alpha,n} \cdot \sqrt{n}/2)) \quad (5)$$

given in [1].

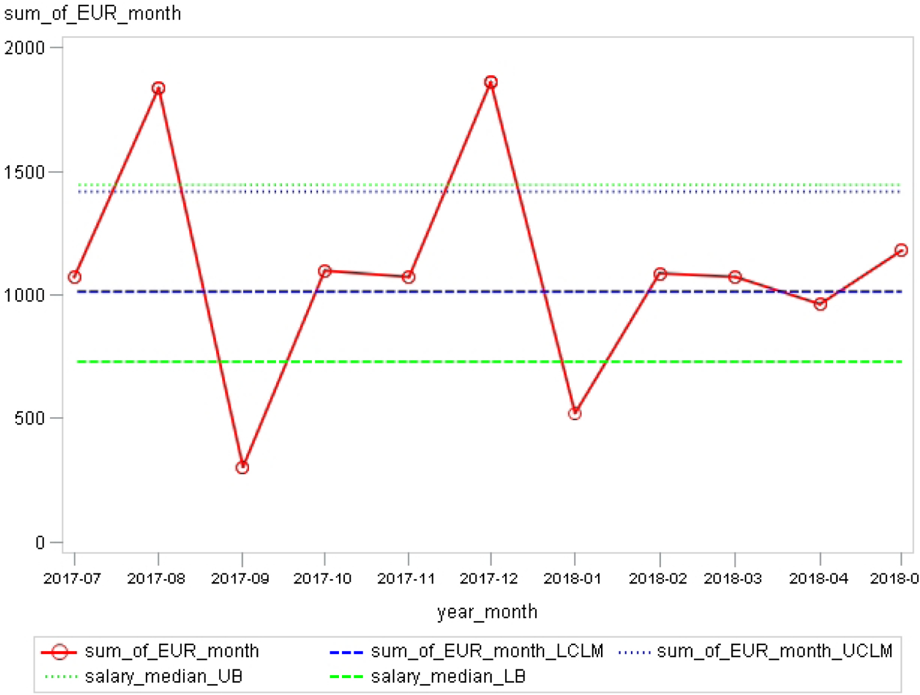


Figure 7: Confidence interval for mean and median salary

This approach may be used to explain expected deviations from median if data are non-normally distributed. Moreover, it is similar to standard confidence interval approach where confidence is understood in sense of magnitude of standard deviations (1-σ correspond to 68%, 2-σ correspond to 95% and 3-σ correspond to 99.7% confidence level).

3.2 A median-based model with quantile factors

One problem with average-based models such as the one proposed by the Bank of Lithuania lies in the dependency of the moving average on outliers. This is of particular relevance for this problem, as it is not untypical in Lithuania to receive two-months worth of salary before an extended vacation.

A more robust approach concerning outliers is to use the median instead of the mean. The building block for the predictive model is thus to use the median of each income stream in the preceding timeframe, i.e. in our case in the preceding year. However, given that not all types of income streams are equally reliant, the medians from the different income streams have to be weighted. This weighting can be done either using a local perspective, that is, assessing the reliability of an income stream on an individual's basis, or using a global perspective, where some reliability factor is computed using the information for the entire dataset.

While the approach of using the local perspective is fairer to the individual, it suffers from the lack of data described in Section 2 and is thus not feasible for this problem. The global perspective has the advantage that it is more descriptive a-priori, that is, it is easier to reason about safety margins etc.

Thus, we propose a model to predict the yearly average sustainable income Y_j of person j as

$$Y_j = \sum_{i=1}^{n} q(i,r) f_{\text{median}} \left(x_{j,i}^{(t)}, x_{j,i}^{t-1}, \dots, x_{j,i}^{t-11} \right), \tag{6}$$

where $i = 1, \dots, n$ iterates over the different income streams and f_{median} determines the median. The key ingredient in the model is the quantile factor $q(i,r)$, a global parameter that only depends on the income stream i and the reliability ratio r. The quantile factor $q = q(i,r)$ is the largest $q \leq 0$ for each income stream i such that $q * \text{median}(\text{x}_{\text{ij}}) \leq \text{mean}(\text{x}_{\text{i}})$ for at least $r * 100\%$ of the people in the dataset. The computation of the quantile factor can be visualised by showing the cumulative distribution over ratios, see figure 8:

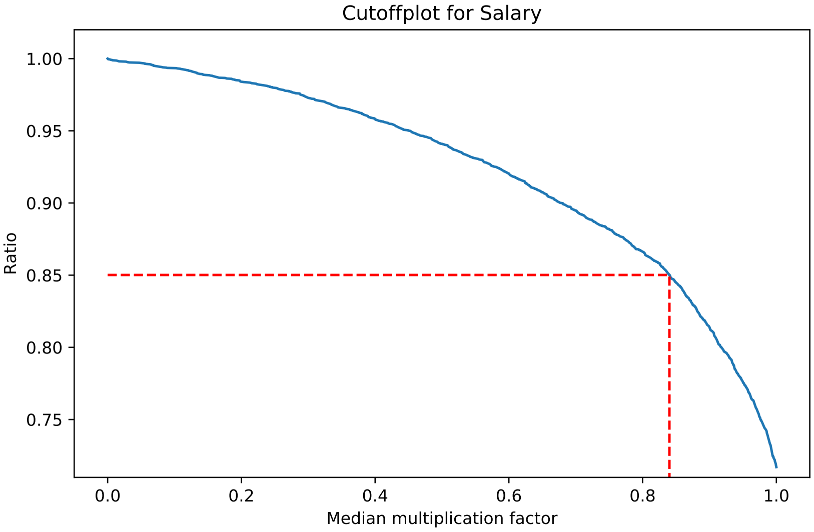


Figure 8: The median multiplication factor $q(i,r)$ for the salary stream (i = salary here) shown with the ratio r which corresponds to including $r * 100\%$ individuals in the dataset in the calculation.

A key question when computing this factor is how to include people who have no information or no income in those income streams. As including people who have no income in those streams will increase q as for the majority of the incomes, the majority has no income in said streams, we only include account holders that have an income in the streams in at least one year and have data for the other year.

The computation can be done in $O(n \log(n))$, but can be sped up to $O(n)$ by using a bucket-sort if only a certain accuracy is required. This can be particularly helpful for large datasets or distributed computing, where sorting the data becomes impractical.

After having determined the parameters, we can gauge the effectiveness of the predictor (6) using an accuracy scatterplot, see Figure 9. The accuracy can be inferred from the difference distribution of the predictor, see figure 10.

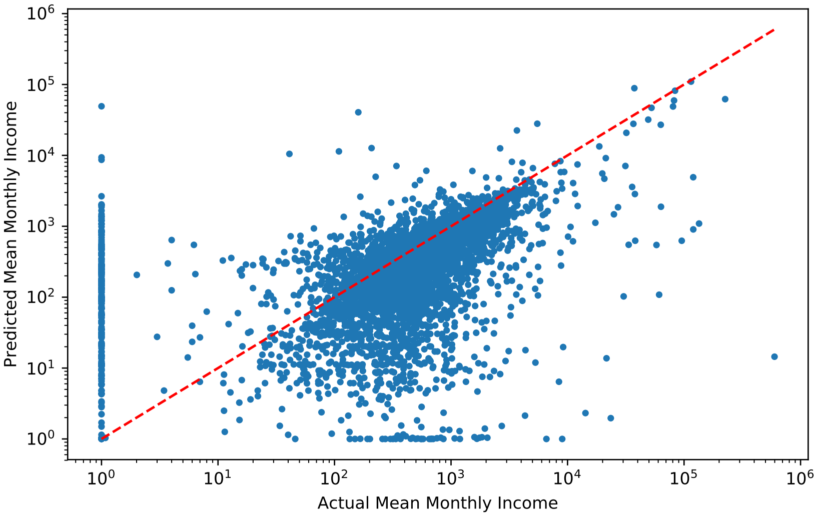


Figure 9: Performance of the median-based model shown by comparing the actual and predicted mean monthly incomes.

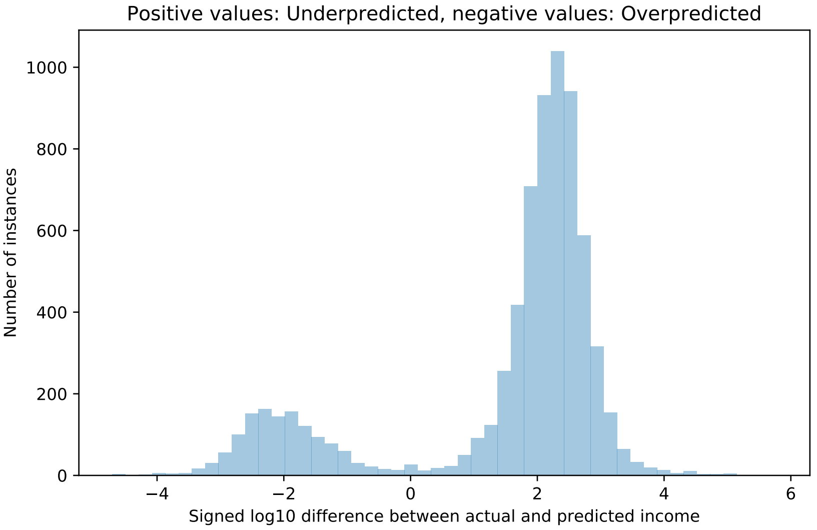


Figure 10: Performance of the median-based model shown by the number of instances of the calculated distance between the actual and predicted incomes.

3.3 Auto-regressive models

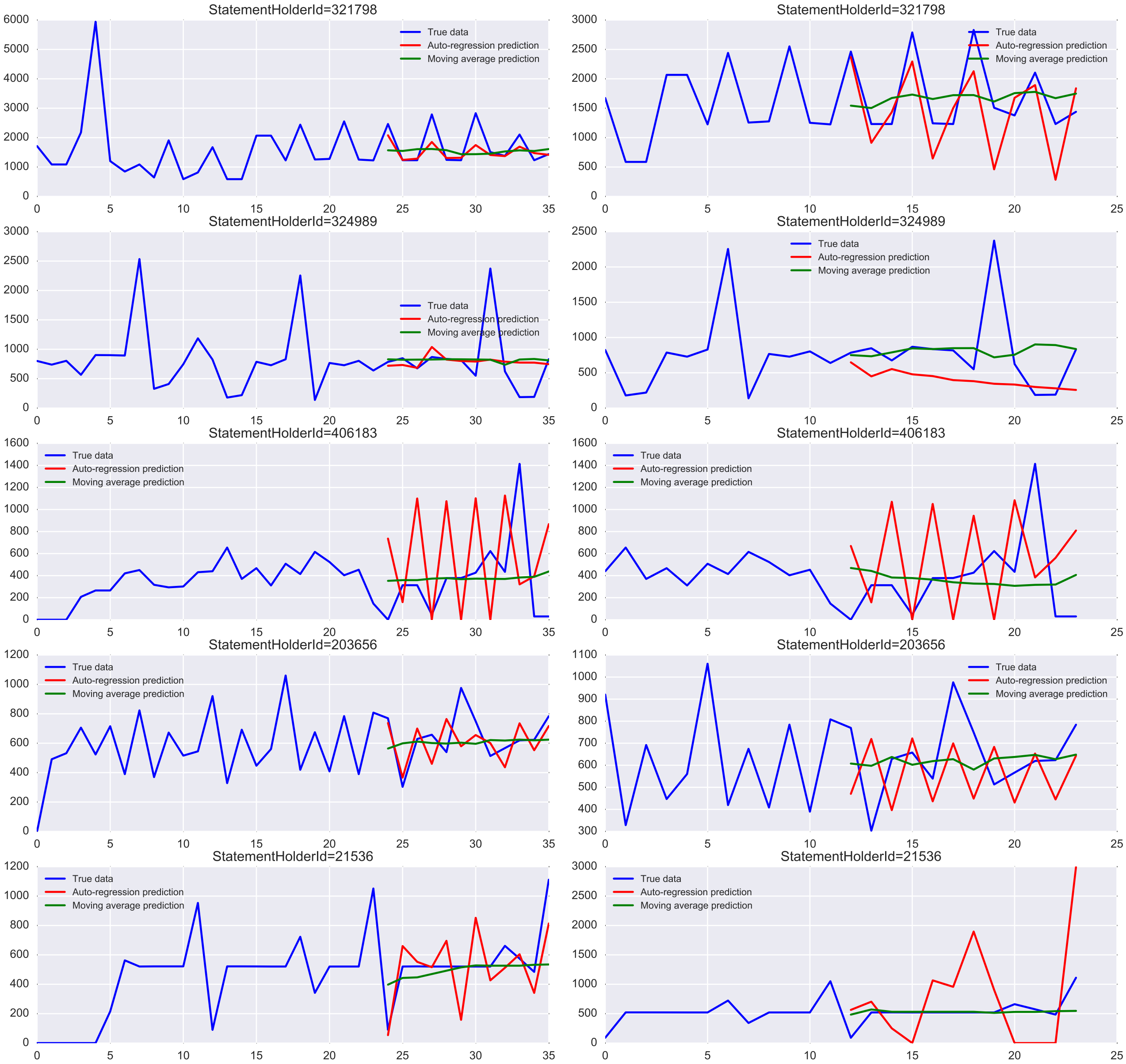
The X_t at time t for one particular customer is given by:

$$X_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t, \tag{7}$$

where $X_{t_1}, X_{t-2}, \ldots, X_{t-p}$ are the values at the previous (known) p steps, and $\{\varphi_i\}_{i=0}^{p}$ are the parameters of the model and ϵ_t is white noise.

In this approach, we fit the model to the past monthly income in one category (for example salary) for one client, and we use the fitted model to forecast the monthly income for the next 12 months. Then we average the 12 predictions to obtain the monthly average income.

In Figure 11, we show a couple of examples where we fit the parameters based on the salary data from the previous 24 (Figure 11a) and 12 months (Figure 11b) respectively and predict the salary for following 12 months (all of the year 2017). We also show here a moving average obtained using a window of the past 12 months.



(a) Using data from the past 24 months.

(b) Using data from the past 12 months.

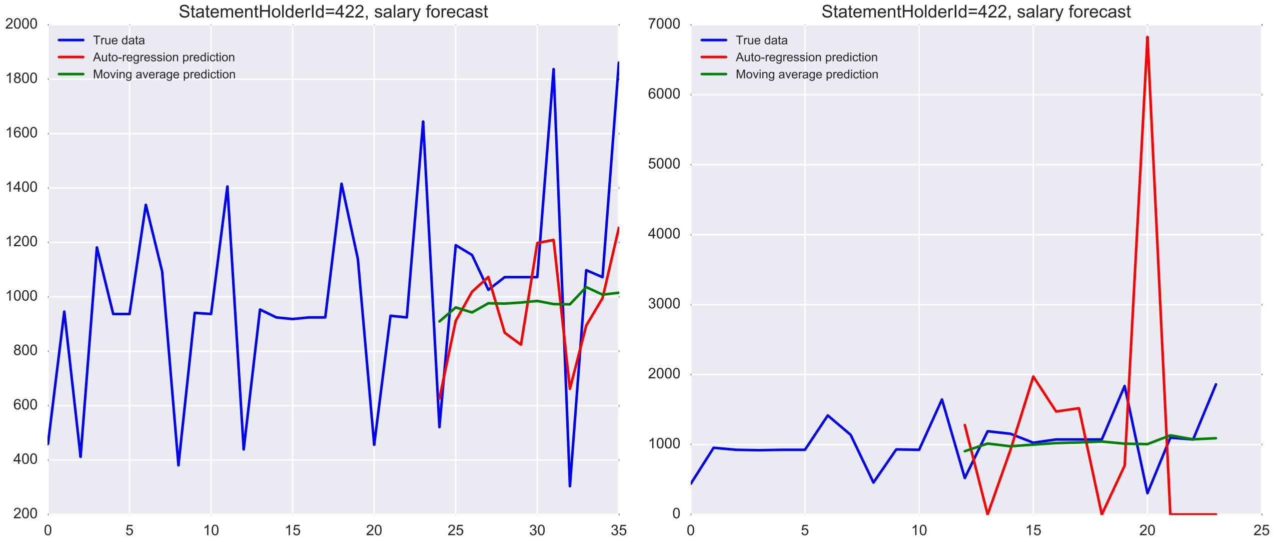
Figure 11: Forecast of 12 months from the past 24 (left) and 12 (right) months using auto-regressive model.

In Table 4, we show the monthly average salary prediction and actual monthly average salary for 2017 corresponding to the monthly predictions from Figure 11. In the column 'Monthly salary' we show the average salary for year 2017 using the real data. In the columns 'AR 24' and 'AR 12' we show the monthly average salary for 2017 obtained from the monthly predictions from Figure 11 using the auto-regressive model trained with the data from the last 24 months and 12 months respectively. Lastly in the columns 'MA 24' and 'MA 12', we show the monthly average salary for 2017 obtained from the monthly predictions using a moving average with a window of 24 and 12 months respectively.

StatementHolderId	Monthly salary avg	AR 24	AR 12	MA 24	MA 12
321798	1723.26	1515.20	1453.20	1540.36	1676.87
324989	798.91	791.59	406.26	818.33	813.55
406183	365.52	573.54	560.74	376.11	364.92
203656	644.21	608.64	562.35	606.73	621.93
21536	547.33	515.60	778.26	495.00	531.63

Table 4: Monthly average salary predicted using the data from Figure 11.

By looking at these examples, we can form an intuition for when the auto-regressive model works best. One advantage if using it is its ability to detect patterns in the data. For example, for StatementHolderId=321798 and StatementHolderId=203656, when the model is fitted using the past 24 months, the spikes in the data are detected, and to a less extent we can see this for StatementHolderId=21536. On the other hand, no pattern is detected for StatementHolderId=406183, in which case the predictions oscillate around the average value. This oscillatory behavious is also seen in most of the predictions done using only the last 12 months, where the only example where the correct pattern is detected is StatementHolderId=321798. This suggests that auto-regression works best when more data is used (here, 24 months). In all these examples, the moving average model smooths the oscillations and follows the general trend of the series.



(a) Using data from the past 24 months.

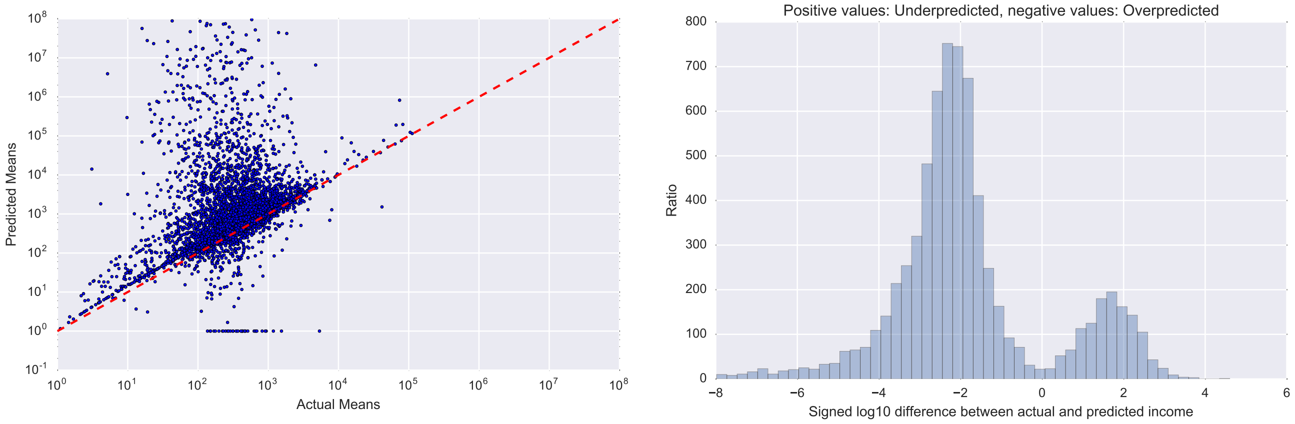
(b) Using data from the past 12 months.

Figure 12: Forecast of 12 months from the past 24 (a) and 12 (b) months using auto-regressive model for StatementHolderId=422.

StatementHolderId	Monthly salary avg	AR 24	AR 12	MA 24	MA 12
422	1106.71	961.26	1224.20	977.81	1024.90

Table 5: Monthly average salary predicted using the data from Figure 12 for StatementHolderId=422.

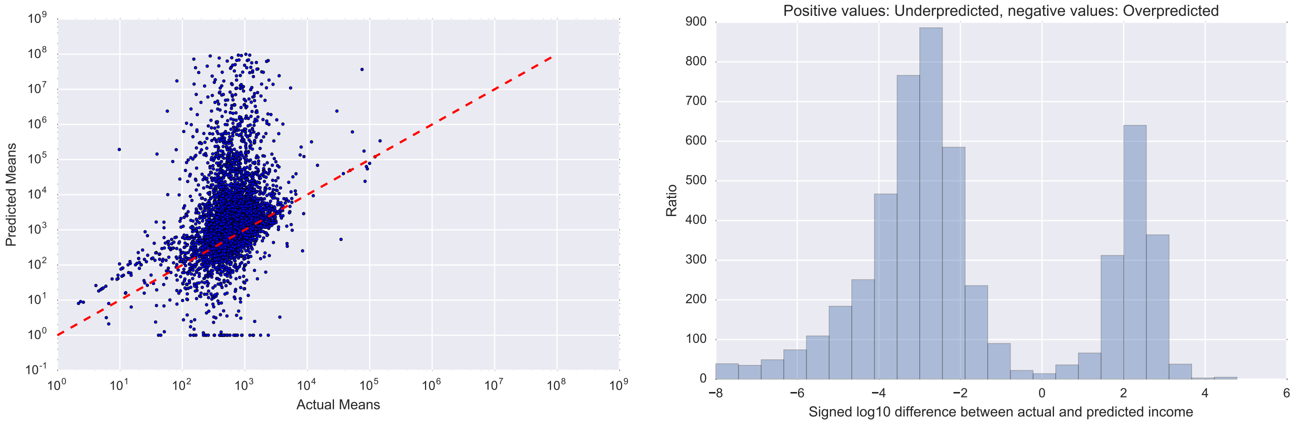
Lastly, we show in Figure 12 and Table 5 the monthly predictions and the predicted average monthly salary for StatementHolderId=422, the same example as in Section 3.1. As in the previous examples, the auto-regressive model works better when using data from the last 24 months, in which case it detects the correct oscillations, but not the full amplitude. While the moving average does not detect the oscillations, it gives the closest prediction to the true average salary, specifically when using the last 12 months as the moving window.



(a)

(b)

Figure 13: We use an auto-regressive model trained using the past 24 months to predict the mean income in the next 12 months.



(a)

(b)

Figure 14: We use an auto-regressive model trained using the past 12 months to predict the mean income in the next 12 months.

In Figures 13 and 14, we applied the auto-regressive model to all the clients 'Salary' data to predict the monthly salary for 2017 using the previous 24 and 12 months respectively. Because some of the data is missing, we back-filled these entries by setting them to be equal to the previous ones. We then applied the auto-regressive model averaged the monthly outcomes to obtain an average monthly estimated salary, after zeroing the negative predictions. Here we are not showing the predictions that are above 10^8. This can happen when we predict too far into the future given the available data. In this case, the data should be evaluated by a human observer, or using a different model. We can see in these plots that the auto-regressive model tends to over-predict, this being more obvious when we only use the data from the previous 12 months. For the 12 months model, both the under-predictions and the over-predictions happen to a greater extend, as the peaks in Figure 14 (b) are further from zero compared to the other models.

4 Conclusion

In this Study Group we discussed a number of approaches to develop an algorithm for calculating sustainable income. While standard practice advised by the Bank of Lithuania (credit market regulator) is to calculate sustainable income from mean salary or pension (averaging over 4, 6 or 12 months depending on type of credit line) only, in this study we have shown that other regular income including social security benefits and bills can also be treated as sustainable income. Data analysis showed that while there are 15 different income streams, for a large majority of individuals only a few will be relevant. Clustering analysis didn't show any significant difference between classes of the clients.

Models based on weighted median and autoregression were proposed for increasing importance of income regularity. Auto-regression model fits better when more data is used (24 months). Analysis of model forecasts based on auto-regression and moving average, showed that in average over-prediction (model says that person will have bigger income than actually is) of sustainable income is bigger than under-prediction (model says that person will have smaller income than actually is). However, if income sources with rare occurrences (frequency is less than 4) are removed from the model then deviations are much smaller.

Further approaches to consider would be to compare a weighted median-based model to a weight mean-based model to test the hypothesis of the importance of excluding outliers in the prediction. Moreover, further algorithms including tree-based regression should be considered for this dataset to provide forecasts of income.

The work undertaken in this week provides the basis for further development of an algorithm to predict sustainable income. The exploratory data analysis highlighted relevant income streams to include, and an initial modelling framework provides a baseline approach to forecast future sustainable income.

References

[1] D J Olive. A simple confidence interval for the median. Preprint, available from http://lagrange.math.siu.edu/Olive/ppmedci.pdf, 2005.

[2] SAS Software. http://support.sas.com/documentation/cdl/en/qcug/63964/HTML/default/viewer.htm#qcug_functions_sect016.htm, 2018.