Estimating Non-Regular Earnings for Micro Organizations: A Microdata Approach

Gergely Attila Kiss[1], Beáta Horváth[2], István Balázs[1]

[1]Hungarian Central Statistical Office, Hungary

[2]Hungairan National Bank, Hungary

Abstract

To satisfy the demands and needs of modern users of statistical products, the paper presents a new method for estimating non-regular earnings for micro organizations in Hungary. The method uses sources of the National Tax Authority and Hungarian State Treasury on income taxes that allow analysis of individual earnings on a monthly level. The core of the method is predicting outliers by combining machine learning methods and subject matter expert knowledge that is extended with sample adjustments by weighing. The promising results of the new method and the ease of generalization to national level makes the method to be a good candidate for creating a consistent method for estimating non-regular earnings on the national level.

Keywords: Machine Learning, Outlier Detection, Panel Data

1 Introduction

This paper presents a new method for estimating non-regular earnings for organizations with less than 5 employees in Hungary. The non-regular earnings statistics are a part of income statistics, they are the difference between the gross and regular gross income. By definition, the non-regular earnings include the premium, salary for the 13th month, and other rewards (Hungarian Central Statistical Office, 2024). It is a high-priority statistic because of the needs of researchers and decision makers, as well as from the public domain. Thus, there are increasing demands on its accuracy, timeliness, and coverage.

There are two difficulties in calculating these statistics. The first is stemming in that these are the smallest functioning economic actors so they are less compliant to spend their resources to create statistics. Creating a traditional data collection-based method for estimating the non-regular payouts would require tremendous effort from both the regulatory and the actor sides.

The second is coming from the fact that in Hungary there is no difference in taxation between the regular and non-regular segment of the income. Thus, even if the Hungarian Central Statistical Office (HCSO) has administrative sources where earnings are observed (which is the case), there is no incentive for the people or the regulatory institutions to enforce the distinction between the two types of earnings. Thus, it is mostly random if an accountant labels a non-regular income as non-regular earning. These difficulties lead me to analyze the change in incomes in the time dimension for the estimation.

The method we present satisfies most of the quality demands while also attaining aggregations of new dimensions due to the appropriate use of administrative micro-data. The novelty of my approach is in mixing traditional and modern solutions for estimating the non-regular earnings for each month since 2019 in Hungary. Furthermore, it will be the first development conducted inside the HCSO that combines traditional and machine learning techniques to produce official statistics. The current procedure at the HCSO to estimate this statistic is based on a macro-level projection of the non-regular payouts of large and small organizations (more than 5 employees) to the target population of micro organizations (less than 5 employees).

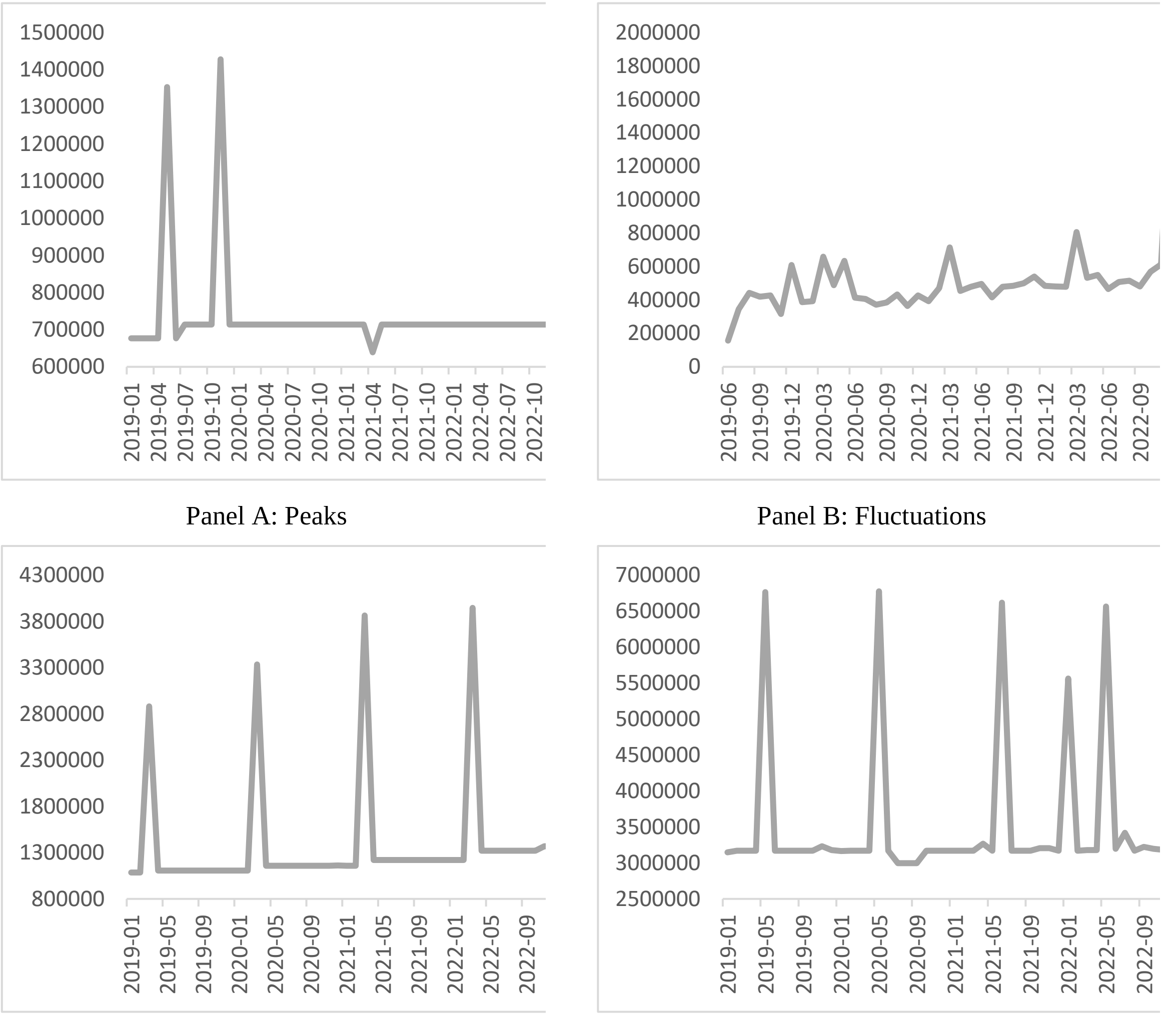
The rest of the paper is going as follows. The second section describes the data, as well as some of its limitations. The third section discusses the outlier detection techniques. The fourth section presents a solution on how to create the final estimates from the analysis sample. The fifth section shows the process of estimation. The last section concludes with the results the need for further testing, and how to generalize the method to the national level.

2 Data Characteristics and Panel Structure

The data originates from the National Tax Authority and the Hungarian State Treasury Hungarian Central Statistical Office - Metainformation 2024, and are on the personal income declarations of employees. The former covers enterprises and non-profit organizations, while the latter provides data on governmental bodies. In Hungary, the monthly declarations are created by the employers' payroll teams, and as such the data records the exact gross payment of the employees. The two sources cover the whole employed population since 2019 Hungarian Central Statistical Office - Metainformation 2024, altogether about 4.7 million distinct individuals.

The panel structure is created by merging the four years of cross-sectional data with a unique identifier that is created from the 4-digit ISCO code, the personal tax identifier, and the organization tax identifier. These three identifiers ensure that our panel only includes data for people who work in the same occupation and at the same organization. The reason to be so limiting in this merging process is due to the uniqueness of non-regular earnings. To measure it we have to filter out the changes in earnings that are due to both sides of the employment relations. Nonetheless, the panel sample still contains circa 1.6 million observations which is still a huge sample considering it is about one-third of the whole cross-sectional data, that is to cover the whole employed population.

Figure 1: Typical shapes of earnings time series



Panel A: Peaks

Panel B: Fluctuations

Panel C: Peaks with level-shift

Panel D: Peaks with fluctuations

The result of this strict merging is that the complications that would come from differences in earnings from promotions or changing jobs are filtered out. For example, if a junior software developer changes positions after a few years (be it inside or outside the organization) and has a jump in salary that would cause a large step or in the case of a signing bonus even a peak in the time series. These kinds of increases could be easily identified as some non-regular payments, especially if it happens too early or late in the analysed period.

One has to note that the time series data that make up the panel sample is not the usual kind that comes to mind. As can see in Figure 1 the time series consists mostly of flat basins in a step case like shape with occasional peaks in them. This atypical shape is due to the economic rigidity coming from the fact that the employment contracts are made for either an indefinite time or for at least a year in Hungary and also usually include at least some fixed wage part. However, this rigidity makes it easier to find the peaks as they are very different compared to other points in the time series.

Finally, to create the final estimates the income numbers are also extended with descriptive variables for both the employee and the employer. These variables include the Hungarian version of NACE and ISCO, the gender of the employee, age categories of the employees, their education, size category of the employer, and type of employer organization. The most important ones to adjust our sample are the age categories, NACE and ISCO both in 4-digit breakdowns. The economic intuition behind using these covariates is that young people do not tend to keep their jobs for long periods such as a 4-year interval. In the case of NACE and ISCO, we expected to catch the industrial and educational differences between occupations and employer types. For example, we would expect workers with low-education to change between jobs more frequently than vocationally or high educated employees.

The descriptive evidence also suggests that there is significant heterogeneity between the panel and pooled cross-sectional data in the categories mentioned above. Table 1 shows that in our panel sample, there is a much smaller proportion of young people, and the panel creation process overweighs the two categories for the eldest. Table 2 provides similar evidence in the case of the category of Elementary Occupations, although that table shows a smaller difference between other categories. These evidences also support the traditional non-response adjustment as these covariates have different distributions inside and outside of the panel, ergo they also have some predictive power of being in the panel sample.

Table 1: Age category Heterogeneity table

Age categories	Panel	Year 2019	Year 2020	Year 2021	Year 2022
< 25	0.5%	14.8%	10.4%	10.4%	10.7%
25-35	12.0%	22.7%	22.3%	22.3%	21.9%
35-45	25.9%	28.3%	26.9%	25.7%	24.7%
45-55	37.1%	22.6%	25.1%	26.2%	26.7%
55-65	23.3%	10.9%	14.0%	14.0%	14.3%
65 <	1.3%	0.6%	1.24%	1.5%	1.8%

Table 2: ISCO class heterogeneity table

ISCO categories	Panel	Year 2019	Year 2020	Year 2021	Year 2022
C0	0.0%	1.0%	0.7%	0.8%	0.0%
C1	10.8%	6.9%	7.1%	7.0%	7.4%
C2	18.0%	14.7%	15.1%	15.8%	15.5%
C3	16.4%	15.2%	15.5%	16.4%	15.7%
C4	6.6%	7.1%	7.0%	6.9%	7.2%
C5	10.7%	10.7%	10.7%	10.4%	10.6%
C6	0.8%	0.7%	0.7%	0.7%	0.7%
C7	12.4%	9.3%	9.2%	8.7%	8.7%
C8	14.6%	13.2%	13.0%	12.7%	12.9%
C9	9.5%	21.1%	21.0%	20.4%	21.2%

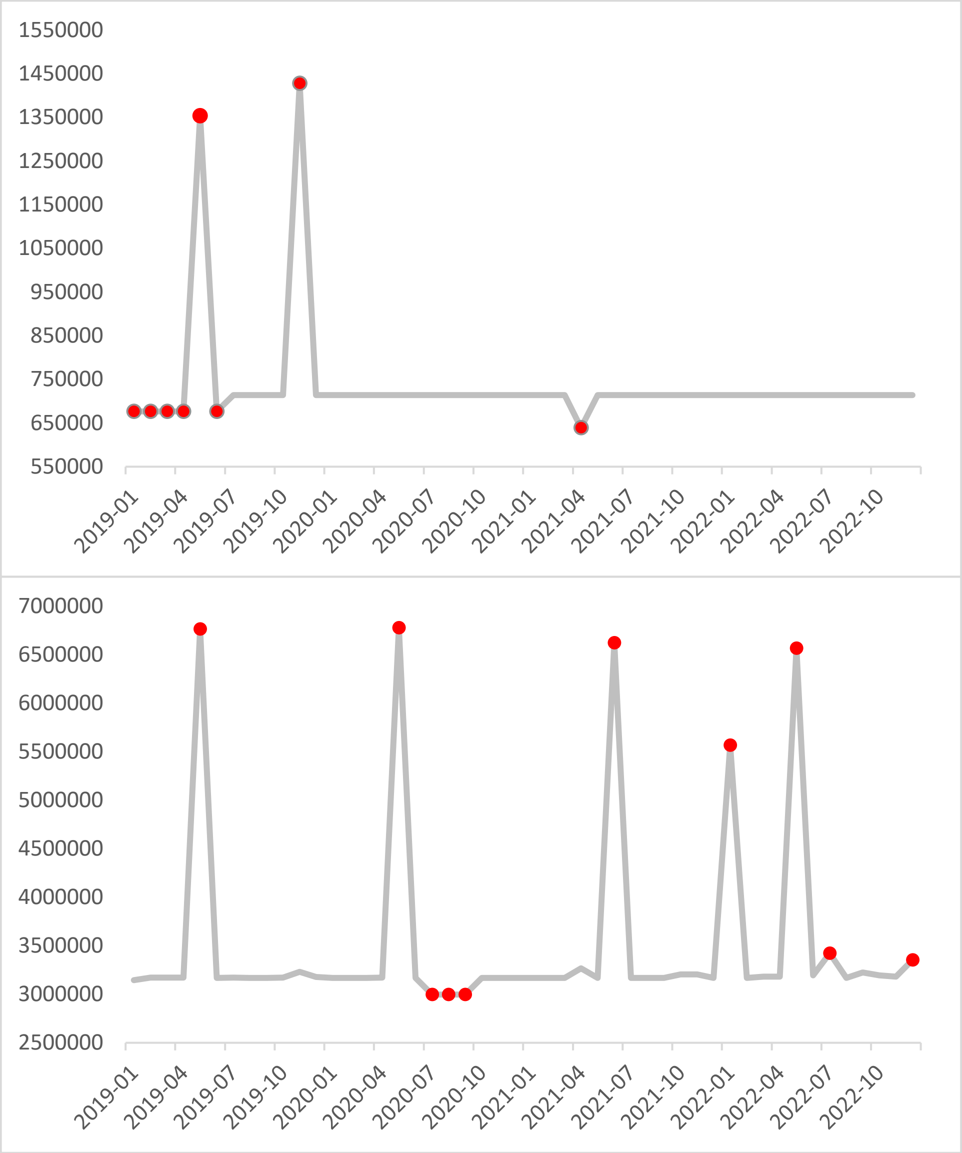
3 Outlier Detection Procedure

The most influential part of the estimation process is the outlier detection. It is a composition of multiple outlier detection approaches we experimented with during the research process. This is extended after several discussions with subject matter experts with rules that heavily lean on their knowledge. The detection process reflects this twofold approach and has two stages too. The first part is the mechanical detection that flags all candidates of outliers. The second part uses economic and regulatory knowledge to implement further rules to approve the flags created in the first part.

3.1 First Stage of Outlier Detection

During the experimentation with the first stage of outlier detection, we tested several different types of detection methods. We tested two rules of thumb based on median earnings inspired by H. Liu, Shah, and Jiang (2004), an ARIMA model-based detection, and two machine learning algorithms, namely isolation forest (F. T. Liu, Ting, and Zhou, 2008) and local outlier factor (Breunig et al. 2000). It is important to note that the first best scenario would have been to use supervised learning for finding the outliers, although there is no such labelled dataset. Therefore, the use of unsupervised clustering mechanisms was the only possible machine learning approach. The most consistent algorithm for finding the possible outliers was isolation forest. The results of the first stage outlier detection can be seen in Figure 2. It is clear that the first stage, in most cases does not make a perfect job. However, it is not expected to be a flawless detector, more like it should catch a large enough base set of possible non-regular payouts that can be further polished in the second stage.

Figure 2: Results of First Stage outlier detection



3.2 Second Stage of Outlier Detection

In the second stage, additional rules are introduced to filter out points from the first stage's results. Starting with the obvious the negative changes are dropped out because a decrease in income cannot be non-regular earning by definition. These rules are formulated based on subject matter expert knowledge that can grab the economic insight and regulatory changes behind events where massive payouts happen. These patterns include retroactive raises, regulatory one-time payouts on a specific date, and how to deal with small or regular fluctuations in earnings.

3.2.1 *Retroactive Raises*

A good example of a retroactive raise was in March of 2021. When employees in public healthcare received a salary increase because of the efforts made in the COVID pandemic, all due to regulation. It was retroactive because the raise should have been active from January on. However, the government only started the payments in March. This means that in March the employees did not just get the raised wages, but also received the difference between the increased wage and the previous wage for the previous two months. This would seem like a massive non-regular payout for any person working in healthcare in March. However, these kinds of payouts should not be included in the target of outliers as actually, all this payment is part of a raise. Even if the previously described example is specific, due to the regulatory factor, this pattern of retroactive salary increases in Hungary is common. The intuition of how to filter out such raises is based on the example above. If there is an observed peak in the series at month mt and the difference between this peak's earnings (w_t) and the average of the previous ***n*** months' wages ($\overline{w}_{t-1,t-n}$) can be divided up to the number of months spent in the year (***m_t***) and this way the difference is not larger than a given ratio (***p***) of the previous n months' average wage ($\overline{w}_{t-1,t-n}$) than it is a retroactive raise (***D_{RR}*** = 1). Formulated as:

$$D_{RR} = \left(\frac{w_t - \overline{w}_{t-1,t-n}}{m_t * \overline{w}_{t-1,t-n}} < p \right)$$

To be able to use this formula there are two decisions to make on *p* and *n*. The former is responsible for the strictness of the filter the larger it is the higher the spikes it demands to consider the peak as non-regular earnings. The second is the number of months to include in averaging the wages. During the outlier detection process, we used *n* = 6 and *p* = .25 after some manual grid search these simple choices proved to be able to generate proper results.

3.2.2 *Regulatory One-Time Payouts*

These cases are dealt with in the simplest way it is possible. If the date, compensation, and subgroup condition match the regulation then the payment is due to it. Since regulatory changes are not predictable by statistical methods there is no other way to detect these types of earnings. One would argue that this kind of payment should be included in the non-regular earnings by common sense, although the regulation can be created in such a manner to be excluded from the statistical definition. Fortunately, there is only one scenario where we have to use it, but it is expected. The scenario is also related to the COVID pandemic, in the period between June 2020 and March 2021 the directors of healthcare-related organizations received a fixed amount of payment.

3.2.3 Small and Regular Fluctuations

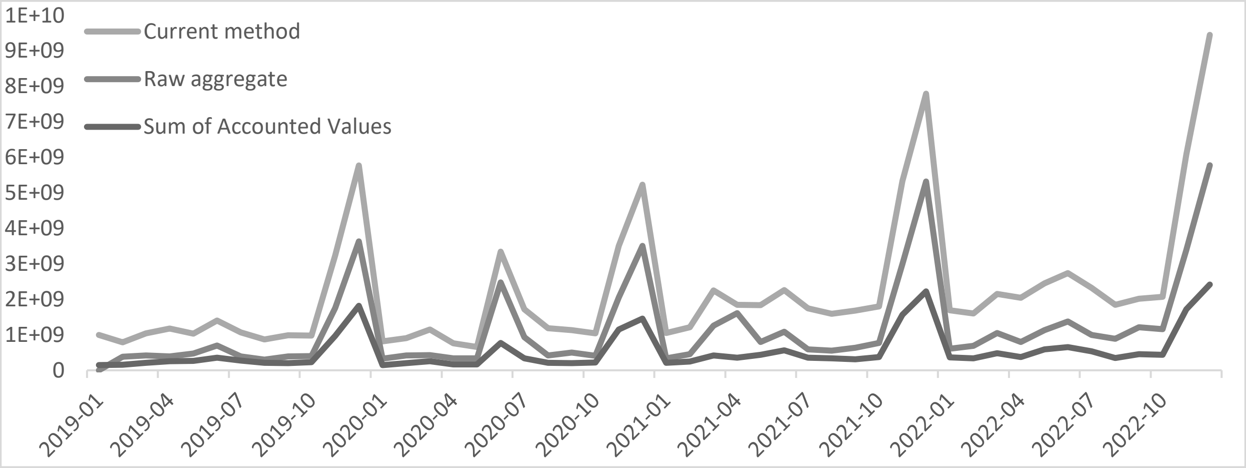
Some small and rather regular fluctuation patterns can be observed in Figure 1. These fluctuations in general could be due to performance-related payment, therefore the small fluctuations under an absolute value are not considered to be non-regular earnings. During the experimentation with this filter, we also tried relative income-based filtering, but that seemed to be much more difficult to use as a general solution due to the large variation in the panel sample in the wages. In the current version, we used the absolute value of HUF 25000 which is around EUR 70-80 in general.

A specific case in these fluctuation patterns is the January relatively larger spikes. These are assumed to be due to renewed definite time contracts and renegotiated wages in indefinite contracts. In the former case, it is usual practice in Hungary that the definite contracts expire by the end of the year, therefore the renewed versions are taking effect by January. In the latter case, renegotiation also takes place at the end of the year causing to have the same timing to affect the earnings. This case dealt with a similar absolute wage limitation, just with a higher bound of HUF 100,000 (EUR 230-250). A further development in both cases is to correct the starting value for each year with the inflation to also cover the possible inflation in the fluctuations.

3.3 Raw Aggregates

To conclude this section, we present some raw aggregates that are the results of the outlier detection procedure. The comparisons of aggregated series can be seen in Figure 3. The different series are: the current statistics from macro-projection, the aggregated values of non-regular earnings after outlier detection, and the sum of values originally accounted as non-regular earnings. The first thing that is clear from the figure is that the dynamics of the time series match. That is already a good sign that our outlier detection process works fine and does not modify the dynamics of the series.

Figure 3: Comparison of aggregate series by occupation types



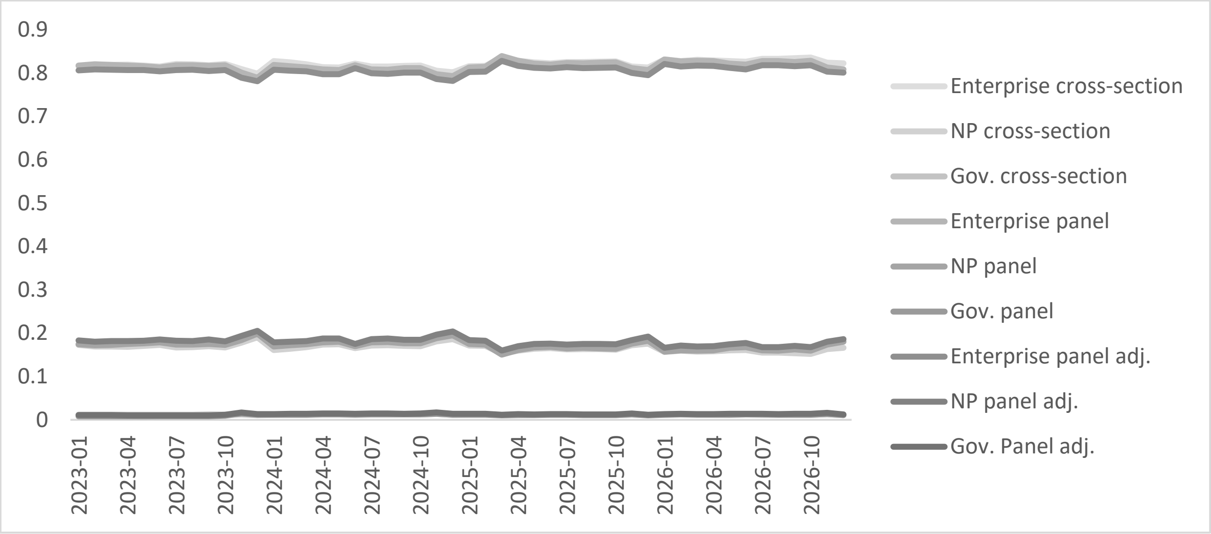
4 Sample Adjustments for Final Estimation

The final step in the estimation is to turn the raw aggregates into one that represents the national level of organizations. To conduct this, a traditional and straightforward approach is to do a weighting and calibration similar to a non-response adjustment (Gary, 2007). The adjustment requires understanding where are the differences between the sample and the target populations, which can be tested with heterogeneity analysis. After understanding the differences, a simple inverse-probability weighting should correct the distributions.

First of all, as it was more of a when and not if question to implement a similar correction for the national level we decided to correct on it. Then we do not have to change the method if we construct an estimator for all organizations while it is still easy to get all the micro ones by a simple conditional query.

A simple heterogeneity analysis shows that the main differences are in the subgroups of young and physical employees. The Tables 1, and 2 already show these patterns. To complete the statistics our target was to create weights to adjust the two currently most important categories for publication. These are categories about the type of organization (entrepreneurship, non-profit, and governmental body) and the type of job (physical, intellectual, or unknown).

Figure 4: Comparison of aggregated series by occupation types



The inverse-probability weights are calculated by a simple logistic regression which is estimated on the 2019 cross-section data, on the binary variable of being in the panel sample. The variables used for estimation are the ISCO and NACE 4-digit breakdowns, the gender of the employee, and the age category of the employee. Since the panel is constructed with the condition that the individuals are continuously working at the same job and the personal covariates do not change much in 4 years, it should not matter much in which year the equation is estimated. However, as the period covers the COVID pandemic we decided to rather use the years that were not affected by it. This decision is based on when we estimated the weights for each year the weights for 2020 were disrupting the dynamics of the time series as that year was different, especially for the manual labor force.

The results of the adjustment can be seen in Figure 4 shows the unadjusted total earnings, the adjusted total earnings, and the cross-sectional total earnings in the job type subgroups in ratios to the monthly total, and Figure 5 shows the same ones in the organization type subgroups. It is clear that the difference diminished meanwhile, the similarities were kept. This is supporting evidence that the heterogeneity is somewhat corrected by the weights for the target categories. The final step is to calibrate the data to the ground truths we have in the data source, our choice for the cornerstone number was the total earnings in months, this calibration was done with a single divide to ensure that the sum of earnings matched the sum from administrative sources. A further improvement can be to calibrate not just to the total earnings but to the total of employed population in the month.

Figure 5: Comparison of aggregated series by organization types

The conducted heterogeneity test and time series comparisons show little to no difference in the target sample characteristics. While comparing time series we set the requirements to be matching dynamics in totals and by the previously published strata. The minimal requirements for the heterogeneity test were to reproduce the same ratios to total in the cross-sectional and our panel sample. As differences emerged between the panel and cross-section data, we decided to adjust the sample by weighting similar to non-response adjustment.

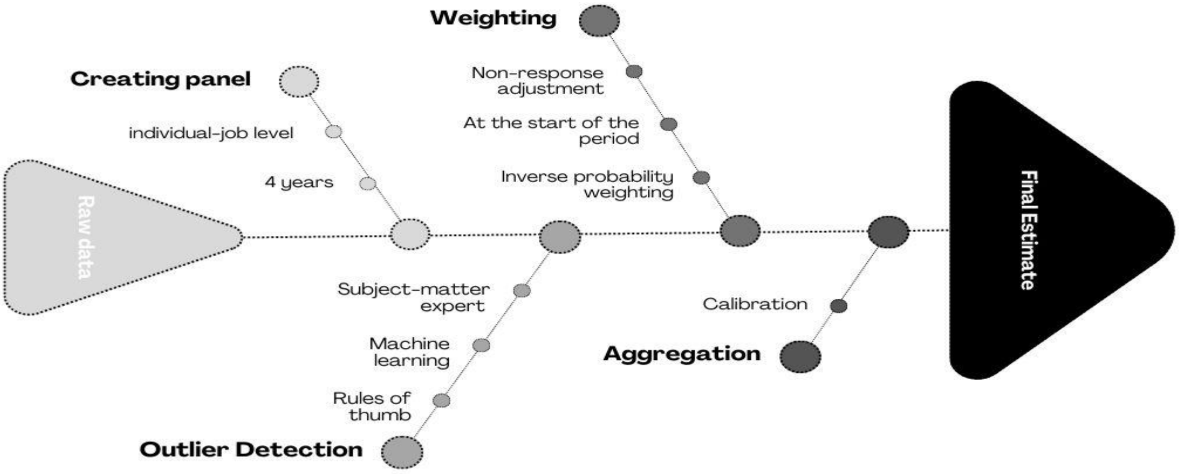
5 Process of Estimtion

The new estimation procedure is a complex solution with multiple stages, which you can see summarized in Figure 6. The backbone of my method is to use the administrative data on income taxes, that HCSO has acquired from the National Tax Authority to its full potential. Therein is data on job-level payouts for each month, so it should be used to construct a presumably more accurate, timely, and detailed estimate of the statistic in question. Therefore, creating an individual-job panel from this cross-sectional job level data is the first step of the process. This version of the data is crucial to be able to analyze when and where the non-regular payouts happen.

The next step in the new method starts with decomposing the panel into individual time series and then running an outlier detection on them individually. Wherein there are easy-to-understand rules of thumb, machine learning classification algorithms, and tweaked results of the former two established by subject matter expert knowledge, detailed in Section 3. The last provides the economic foundation of the filtering process and a deeper understanding of the possible patterns of earnings.

Then, there is a two stage non-response adjustment that includes weighting and calibration. In the first stage, the inverse probability weights are created from the estimated probability of getting into the panel for each individual. This should account for the characteristics-based differences in the strata between the panel and the cross-sectional data. In the second stage, the weights are calibrated to maintain the marginal totals of earnings in the month. This can be done by using a simple ratio of the weights by the earnings to total earnings. After arriving at the final version of the weights a weighted summation provides the final estimates of non-regular earnings.

Figure 6: Flow chart of the estimation process



6 Concluding Remarks

To sum up, the discussed method is a straight approach to finding a solution on how to measure non-regular earnings by simple outlier detection and filtering techniques. It mostly exploits the uniqueness of the individual earnings time series, namely that the earnings show rigidity so that extra payments are easily identifiable as peaks. The difficulties are in how to distinguish between these peaks that match the statistical definition too.

There is a foreseeable issue with the method. If the panel will be regenerated each month as a rolling panel sample, then the models estimated for each individual have to be re-estimated monthly. This means re-estimation for circa 1.6 million time series monthly, or storing that many models and only estimating the newcomers in the sample either way, there are technical requirements to overcome. However, these requirements are not very limiting as the estimation for 1.6 million observations took about a fortnight using 3 standard office computers at HCSO, with a processing speed of about 2 iterations per second.

One effect of implementing this method in production is to provide the non-regular earnings statistics as aggregated numbers based on microdata. Therefore, it will make all kinds of breakdowns attainable that were not possible before. The large coverage of the data and the promising results provide the opportunity to generalize the new method easily to the national level and thus, publish more detailed breakdowns for any organization size. This generalization would make a consistent approach to estimating non-regular earnings for the whole population, as well as providing more detailed numbers for existing publications.

A further outcome could be to replace the survey-based estimates in the long run. The described method does not depend on any survey data so it makes the estimation possible without it. However, we have to be careful with the replacement, as the survey is quarterly. In the short term, it should be just an extension to the survey, that shows more detailed numbers.

There is an additional need for testing before using the method in production. There should be a test to measure how much computational capacity and time it takes to create the estimates. It should be combined with a testing of the process flow to consider how it will work in production and use a test period to create estimates monthly. This is mandatory to be able to plan with the use of the estimation method for official statistics publications.

References

Breunig, Markus et al. (June 2000). “LOF: Identifying Density-Based Local Outliers.” In: vol. 29, pp. 93–104. doi: 10.1145/342009.335388.

Gary, Pike R. (2007). “Adjusting for Nonresponse in Surveys”. In: Higher Education: Handbook of Theory and Research. Ed. by John C. Smart. Dordrecht: Springer Netherlands, pp. 411–449. isbn: 978-1-4020-5666-6. doi: 10.1007/978-1-4020-5666-6_8. url: https://doi.org/10.1007/978- 1-4020-5666-6_8.

Hungarian Central Statistical Office - Metainformation (2024). url: https: //www.ksh.hu/apps/meta.objektum?p_lang=EN&p_menu_id=110&p_ot_id=100&p_obj_id=ABCA (visited on 02/09/2024).

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou (2008). “Isolation Forest”. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. doi:10.1109/ICDM.2008.17.

Liu, Hancong, Sirish Shah, and Wei Jiang (2004). “On-line outlier detection and data cleaning”. In: Computers Chemical Engineering 28.9, pp. 1635–1647. issn:0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2004. 01.009. url: https://www.sciencedirect.com/science/article/pii/ S0098135404000249.