DataSentics

Development of dynamic micro-simulation model of the pension system of The Ministry of Labour and Social Affairs – III.

Project VS/2018/0380 “Development of microsimulation tools for social insurance projection (DEMTOP)” has been funded with support from the European Commission. This feasibility study reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Project VS/2018/0380 Development of microsimulation tools for social insurance projection (DEMTOP) and this document has received financial support from the European Union Programme for Employment and Social Innovation “EaSI” (2014-2020). For further information please consult: http://ec.europa.eu/social/easi The information contained in this feasibility study does not necessarily reflect the official position of the European Commission.

Content

Disclaimer

Loss of functionality can occur if unauthorized changes are made to the software source code. User support is unavailable to address any problems caused by such actions. We recommend not to alter the code, with the exception of updating the .fac tables.

1. Aim and Structure of the Document

This documentation constitutes one of the outputs of public procurement for "Development of dynamic micro-simulation model of the pension system of The Ministry of Labour and Social Affairs – III." The subject of the public procurement contract was the development of model NEMO, i.e., an analysis of provided data sources and an implementation of selected factors in the model according to specification provided in Appendix No. 1 Contracts[1], details described in Feasibility study (Deloitte, 2014) and regular consultations conducted by the client and the service provider regarding further specification of the relevant topics.

The introductory chapters of this documentation briefly summarise the character of implemented changes and used data sources.

The main body of this documentation is divided into individual areas within the scope of work for this public procurement:

1. Concurrence of Payments of Old-Age Pension with Income from Gainful Activity
2. Commencement of Retirement Pension
3. Care for Dependents
5. Region
6. Occupation
7. Wage
8. Self-Employment

This documentation describes outcomes of analyses for each of these seven areas (including processes, data preparation, evaluations of significance of analysed factors and recommendations for implementation) and also the implemented changes themselves (changes in input model point database, the DCS tool and the Prophet model).

Technical documentation comprises a separate appendix, it describes source codes used to create outcomes of this project, such as assumptions tables, graphs or columns which extend the database of model points.

Outcomes of Chapter No. 4 Sickness Rate and Disability are described in a separate document Feasibility Study of Implementation of Additional Factors for the Subject of Sickness Rate and Disability.

Model NEMO with implemented changes was a subject to testing and the results of these tests were communicated to the client; these results are not included in this documentation.

[1] Contract on development of multicriteria decision processes in the micro-simulation pension model of The Ministry of Labour and Social Affairs, as described in Appendix No. 1 from 23rd June 2021

2. Summary

The MPSV dynamic micro-simulation pension model (herein “model NEMO”) is used to simulate individual life paths of a broad sample of the Czech population (so called model points), and thus to predict future developments of the pension system of the Czech Republic (Plívová, 2011). This solution comprises multiple parts; mainly:

- Output database INEP_PARTICIPANTS of so-called model points (further also referred to as MP). Each model point has several attributes such as gender or age and represents one so called main person and potentially also so called secondary persons (for example a husband or wife). This database is in file format .csv and it is an output of a separate solution (public procurement Creation of Analytical Tools I.; its modification is not within the scope of this project). The model points used in the Prophet tool themselves are created from this database with the Prophet Data Conversion System tool (herein DCS).
- The model (workspace) implemented in the Prophet tool. This model describes the logic of the simulation: according to which rules should the characteristics of the individual model points develop during the time steps of the simulation.
- Tables (.fac) of assumptions from which the Prophet model reads for example probabilities of events and transitions from one status to another.
- Tool DCS which creates individual model points for the model. It prepares both the existing persons as well as persons newly born in the future and immigrants.
- Tables (.fac) of assumptions for tool DCS which are used to determine some of the model point characteristics, such as future occupation of new-borns.

The current project affected all of these areas in the following way:

- Based on the data analysis, we proposed and implemented an expansion and adjustments to the logic of the Prophet model and also to the logic of model point creation in DCS.
- We carried out calculations of new .fac tables of assumptions and adjustments to some .fac tables of assumptions for the Prophet model and DCS.
- We created new codebooks used in the model, specifically converters of information about region and occupation to a suitable granularity.
- We created new columns to add to model points, such as for example postal code (PSČ), in the form of CSV files comprising two columns: unique identifier of a person (`ID_OSOBA_AN`) and a new column (for already mentioned postal code (PSČ): `INIT_ZIPCODE`). These new CSV files are integrated into the model points with the DCS tool. These columns were calculated only once, and it will be suitable to recalculate them (for year 2020 and onwards) when new data is obtained. We recommend considering inclusion of an automated calculation of these new columns (model point attributes) directly in the solution for the creation of database INEP_PARTICIPANTS. The calculation process is described in appendix Technical documentation and in the source codes which are parts of the outcomes of this order.

Moreover, we assessed the potential of expansion of the model for new factors in individual subject areas (with respect to availability, quality of data and significance of the factors) and we recommended potential increase of scope of the implementation in the future.

3. Source Data and Their Preparation

3.1. Main Sources

Main inputs for the analyses and calculations are mentioned in the following sections.

3.1.1 STATMIN VZ

Database STATMIN VZ is an aggregate of all persons with submitted Personal records for pension insurance.

3.1.2 Extended STATMIN VZ

Database comprising a subset of Personal records for pension insurance enriched with additional data from the Average salary information system (ISPV), especially information about occupation and education of persons.

3.1.3 STATMIN VZ OSVČ

Database STATMIN VZ OSVČ is an aggregate of all self-employed persons who submitted a Statement of income and expenses.

3.1.4 STATMIN ANOD

Database STATMIN ANOD comprises an aggregate of persons who were awarded old-age pension at the end of calendar years.

3.1.5 INEP

Database INEP represents entries of individual records obtained from historical data of filed claims.

3.1.6 SEE20

Database SEE20 is an aggregate of all records on termination of temporary incapacity for work.

3.1.7 PnP

Database PnP is an aggregate of cases claiming care allowance benefits. The database does not include benefits for years 2012 and 2013 due to a change in database system provider. ID identifiers of caregivers and dependents are not compatible with other databases.

3.1.8 DELNEZ

Connector DELNEZ includes the most probable pairings among individuals' identifiers in SEE20 and STATMIN VZ / STATMIN VZ OSVČ / STATMIN ANOD. This connector was created within the scope of a previous project, specifically "Creation of a database to store identifiers of persons for analysis concerning temporary incapacity for work and disablement" (DataSentics, 2019) and it was extended within the scope of this order.

3.2. Additional Sources

The following data sources were used to compliment the main ones. Only sources which were used across multiple subject areas are described in this section. Data sources specific to individual subject areas are mentioned directly in the chapters for the given subject areas.

3.2.1 Education Codebook

A codebook for education was created to enable, after being connected to database Extended STATMIN VZ, a determination of an achieved level of education for every individual. In the original database, education is expressed with one-letter codes from core educational branches (KKOV) classification in range A - V. For the purposes of this project, these codes were mapped onto four levels of education as described by the Czech Statistical Office – Primary education and none (KKOV codes A—C), Secondary education with a vocational certificate (D—J), Secondary education with completed "maturita" graduation exams (K—M) and University education (N—V).

3.2.2 Occupation Codebook

Database Extended STATMIN VZ includes information about performed occupation in the form of a CZ-ISCO code. To simplify the analyses, codebook of the Czech Statistical Office (ČSÚ) was used (ČSÚ, Klasifikace zaměstnání (CZ-ISCO), 2020).

3.2.3 Postal Code (PSČ) Codebook

Data sources (especially STATMIN VZ) include information about the residence of a given person or about the location of main office of the employer in the form of a postal code (PSČ). However, this information is too detailed for most analyses, and it was necessary to convert it to a district level, potentially with a differentiation between an urban and rural area. Therefore, a conversion table was prepared based on a database of the Czech Post (ČSÚ, Seznam částí obcí a obcí s adresním PSČ, Česká pošta, 2021), this converter assigns each postal code to one district and determines its locality (urban / rural area). This converter was enhanced to include number codes of OSSZ, which are used, inter alia, in database STATMIN VZ OSVČ, and district codes LAU. A detailed process of assigning postal codes is described within Chapter 5 Region, Factors Impacting Occupation 8.4. The final codebook includes postal codes, district and locality, and it is saved in Prophet as table `zipcodes.fac`, see Table 7.7.

3.2.4 Demographics Table

Some analyses require knowledge of the total number of population of a given age and gender in individual years. For these purposes, a table that includes the composition of the population at the end of year in years 2000 to 2020 was created based on data from the Czech Statistical Office[2] - it has four columns: calendar year, age, gender and population size.

[2] Age composition of the population – units of age, Czech Statistical Office

4. Chapter 1, Concurrence of payments of old-age pension with income from gainful activity

4.1. Introduction

Point of interest in this section is the concurrence of old-age pension with gainful activity qualifying for entitlement to a raise of the awarded pension amount (further referred to only as “concurrence”) with the aim of modelling social security system contributions and potential increase of awarded pensions. Such gainful activity is in text further referred to only as “work”.

It can be expected that the frequency of concurrence differs across different groups of population in respect to their situation, education, etc. In this chapter, we introduce an impact analysis of such factors and a possible method of implementation for some of them.

Only old-age pension was analysed, other types of pensions were not considered, and their method of implementation remains unchanged.

4.2. Data Sources

An actively working pensioner can be detected by joining databases STATMIN VZ, STATMIN VZ OSVČ and STATMIN ANOD.

4.3. Detection and Frequency of Concurrence

Here, the term concurrence is used for work qualifying for entitlement to a raise of the awarded pension amount (i.e., work with compulsory payments towards insurance) starting the first day of the next month after that person had gone into retirement. Work performed during the month in which the retirement is started is not included in the calculations as it is most common that work in that month is continued until the end of the calendar month, and so including such a month would result in inaccuracies in the data.

By comparison with information of The Czech Social Security Administration (ČSSZ) regarding the number of applications for a work-related pension increase, it was discovered that the number of working pensioners is significantly higher than the number of applications, which means that not every pensioner takes advantage of their entitlement to a pension raise. Therefore, a probability of an application for a pension raise for reasons of concurrence was added to the Prophet simulation. This probability determines for each model point whether a pension raise will be applied for or not, that is to achieve the best possible assessment of retrospectively increasing one’s pension amount and its simultaneous adjustments. In case of an application for a pension raise, this model point requests the increase always at commencement of entitlement based on number of worked days. In case of the alternative scenario, the model point never requests a raise.

This approach disregards the following two events, however, neither one proved to be significant:

- If a pensioner applies for a pension increase after more than 5 years of concurrence, the retroactive pay is paid out for the last 5 years only. However, with our approach the increase is applied from the very beginning of concurrence, that is for the whole period of concurrence longer than 5 years. This inaccuracy is insignificant given the low frequency of occurrence of such an event.

- The described approach does not take into account situations when a pensioner who has submitted an application for a pension increase continues to work but does not apply for another increase even though they become entitled to it. This error is also insignificant because data analysis shows that people who have applied for a pension increase once will do so in the future again.

4.4. Impact Analysis of Factors

In the Prophet simulation, the state of concurrence is modelled as a product of events described below. Each of these events is controlled by in advance defined probabilities that are dependent on characteristics of model points, so called factors. The probability model, in accordance to the assignment and also to conclusions of the feasibility study, was besides the already existing factor of gender extended to include two new factors, specifically the highest education achieved and a total duration of employment. As described below, records about achieved education are present in the data in lower quality which prevents correctly performed analysis. This factor was implemented in the simulation, however, probabilities were not calculated with it. Therefore, this factor is de facto not taken into account in the simulation.

An analysis of factors of types of employment, marital status and income was also carried out. Data availability for each of the factors is described. In case of good data availability, impact analysis is described. If the impact was evaluated as significant, suggestion for and assessment of its implementation is included.

Taking into consideration the current implementation method of concurrence in the simulation, analysis focused on factor impact on three different probabilities:

1. Probability of continuation of employment at the moment of old-age pension commencement. In compliance with legislation and implementation in the simulation, transitions from early retirement to regular retirement were not considered, neither were commencements of early retirement.
2. Probability of termination of employment while already in old-age pension. Here, we describe annual probabilities; i.e., percentage of people in concurrence who become inactive during a given year.
3. Probability of return to employment after inactivity in old-age pension. This probability was newly added because data showed that the frequency of pensioners' transitions from inactivity to work is not negligible. It is also an annual probability.

Probabilities regarded in Point 1 were calculated as percentage of people who at the moment of commencement of old-age pension stay actively working, always taking into account a given combination of factors.

To calculate probabilities regarded in Point 2, the number of transitions from concurrence to inactivity in a given year is divided by the number of people who were in concurrence in the previous year, again taking into account a combination of factors. For aggregation of percentages for each year, weighted average is used with the weight being the number of available data; i.e., the number of persons in a given combination in a given year.

Lastly, an analogical process was used for probabilities referred to in Point 3. Here, the number of transitions from inactivity to concurrence was divided by the number of people in inactivity.

Probabilities regarding transitions between concurrence and unemployment are not described and were not implemented in the simulation. The reason for this is that a given event either does not make sense or occurs only for a minimum number of cases.

4.4.1 Gender

Availability of Data

Information about gender is available directly in all used databases in the best possible quality.

Factor Impact

When reaching old-age pension, 46% of women and 44% of men continue working. Rates of termination of gainful activity are almost identical for both genders – 23,1% of working men and 23,3% of working women terminate their gainful activity each year. Future returns from inactivity back to work are more frequent for men - 1,7% of inactive men and 1,1% of inactive women return to work each year.

Altogether, this factor is rather at a medium level of significance. Nevertheless, due to its linear availability, there are no barriers to its implementation.

4.4.2 Education

Availability of Data

Information about education is available in good quality directly in the extended STATMIN VZ database; due to the origin of the database, however, it describes only a fraction of the population. In the context of analysis of concurrence of gainful activity and retirement, information about education is available for 34% of relevant persons.

Using such data would lead to significant distortion of results since the nature of the available data indicates that these data are data for working segment of the population where the occurrence of concurrence of work with retirement is naturally more frequent than in the entire population. While 45% of people continue working a month after commencement of their old-age pension in the entire population, this number is 60% when looking at data from extended STATMIN VZ database.

The only other source of data about education is in model point database where the data was generated stochastically. Checking data from the extended STATMIN VZ database shows independence between real and generated values. Furthermore, it holds that in model points distribution of income is equal for all levels of education. This indicates that data was generated in such a manner to illustrate distribution of different levels of education in the population, for our use, however, such data is unusable.

Because of insufficient data and described distortion for data that is available, we currently do not recommend including the factor of education in the simulation. Doing so would result in incorrect results of the simulation. It Is impossible to correctly describe differences among different levels of education with the data currently available, that is why associated probabilities are unavailable.

Factor Impact

Following figures was calculated with the use of data about education from the extended STATMIN VZ database. The higher the education achieved by a person, the more frequent is concurrence of work and old-age pension. While only 51% of persons with primary education continues work after they

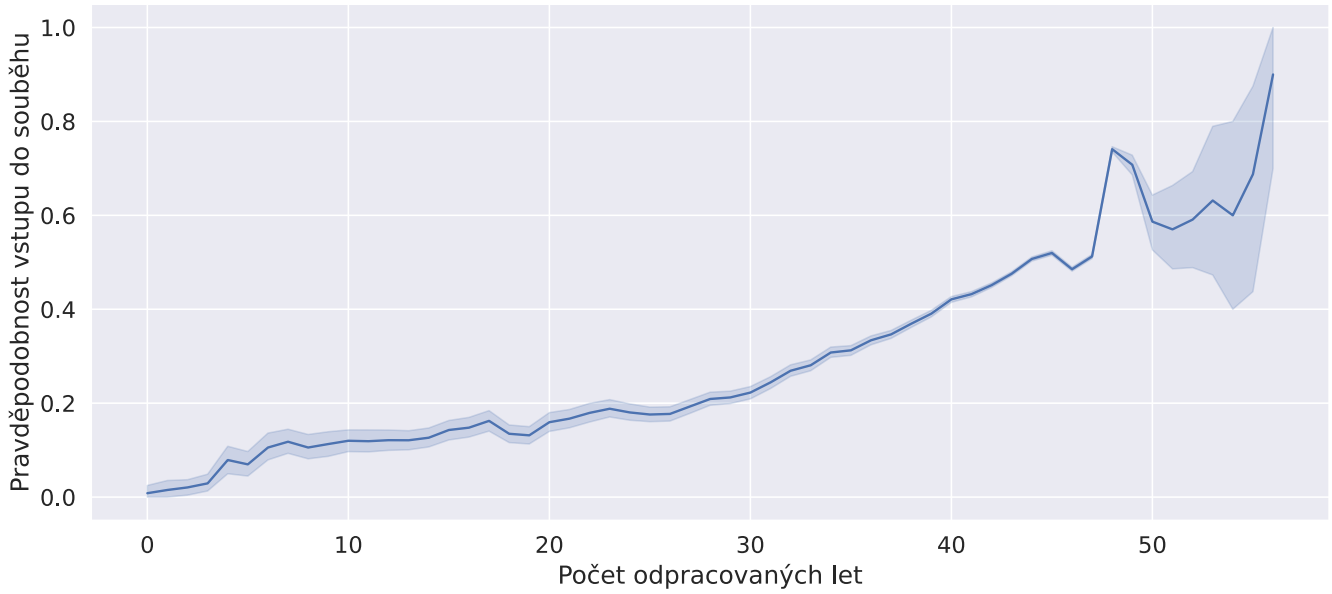

Figure 4.1: Dependence of probability of commencement of concurrence of work with old-age pension at the beginning of retirement on duration of work (expressed in number of years). The curve shows an average value and a 95% confidence interval.

had gone to retirement, this number is 66% for population segment with university education. For secondary school education without having taken and successfully passed "maturita" graduation exams and for secondary school education with having taken and successfully passed "maturita" graduation exams, data shows 54% and 62% respectively.

Analogical tendencies can be observed also when assessing transitions between concurrence and inactivity in pension age. While work in retirement is terminated every year by 23% pensioners with primary education, this number continuously decreases up to only 15% pensioners with university education. Similar tendency remains also for return to work from inactivity where values increase from 10% for persons with primary education to 13% for persons with university education.

This factor shows to be very significant with a strong impact on the development of the observed event.

4.4.3 Total duration of employment

Availability of Data

Information about duration of insurance was used for this factor. Its sum can be easily calculated with the use of INEP database.

Factor Impact

As per expectations, probability of concurrence of work and old-age pension is at the lowest level for the population segment with the least number of worked years, from there it grows continuously, see development of the curve illustrated in Figure 4.1.

Similar curve just with the opposite trend can be seen also for probabilities referring to terminations of work and becoming inactive in pension age. For these reasons, we consider this factor very significant.

This relationship is implicitly influenced by health condition of a person: while there are 35% people who receive some type of disability pension in the group of people who have worked fewer than 35 years, in the group of people who have worked more than 35 years, this percentage is only 16%.

This numerical variable was divided into 7 intervals for easier implementation. First interval includes people who have worked for less than 20 years, the last interval those who have worked for more than 45 years. There are 5-year long intervals between these values.

4.4.4 Type of Employment

Availability of Data

Based on the code of gainful activity stated in Personal records for pension insurance (ELDP), contract types in STATMIN VZ database can be distinguished as employment contract, agreement to perform work (DPČ) and agreement to complete a job (DPP).

With the use of STATMIN VZ OSVČ database, concurrence of work and self-employment can be easily determined. On the contrary, full-time and part-time jobs cannot be distinguished from another using the current databases.

Factor Impact

The lowest probability of concurrence with going to retirement is for agreements to complete a job (DPP) and it is 27%, then for agreements to perform work (DPČ) at 50%. In comparison, this probability for strictly self-employed people without other contractual work is 59% and for people working strictly on employment contract is 57%, thus we see very small difference in probability of concurrence between these two groups. The probability is higher for people who at the time of commencement of their retirement were working under concurrence of multiple types of employment. The highest probability of concurrence is for those who continued working under employment contract and at the same time were self-employed (77%).

This factor is considered less important due to an overall small amount of people working under agreements to perform work (DPČ) and agreements to complete a job (DPP) and prevailing occurrence of employment contracts and self-employment in pre-retirement age.

Similar relationships can be observed also for terminations of concurrence of work with retirement; i.e., when a person becomes inactive by terminating their gainful activity in retirement. This is most frequent for agreements to complete a job (DPP) at 56%, followed by agreements to perform work (DPČ) at 28%. Values for self-employment and employment contracts are almost identical (20% and 21% respectively). The lowest probability is for people with concurrence of multiple employment types.

Here, the factor becomes interesting since agreements to perform work (DPČ) are frequent in pension age.

Inverse relationships exist for return to concurrence of retirement with gainful activity after inactivity in pension age – when pensioners conclude one of the agreement types which are in consequence more frequent among pensioners than in the entire population. The same holds for self-employment which is performed in pension age more often than is standard in the population.

4.4.5 Marital Status

Availability of Data

In STATMIN ANOD database, data about number of children for women is easily available, and so is information on whether a person is widow(er). Due to poorer data quality in regards to number of children while taking widow(er)'s pension, the only records which were used were old-age pension

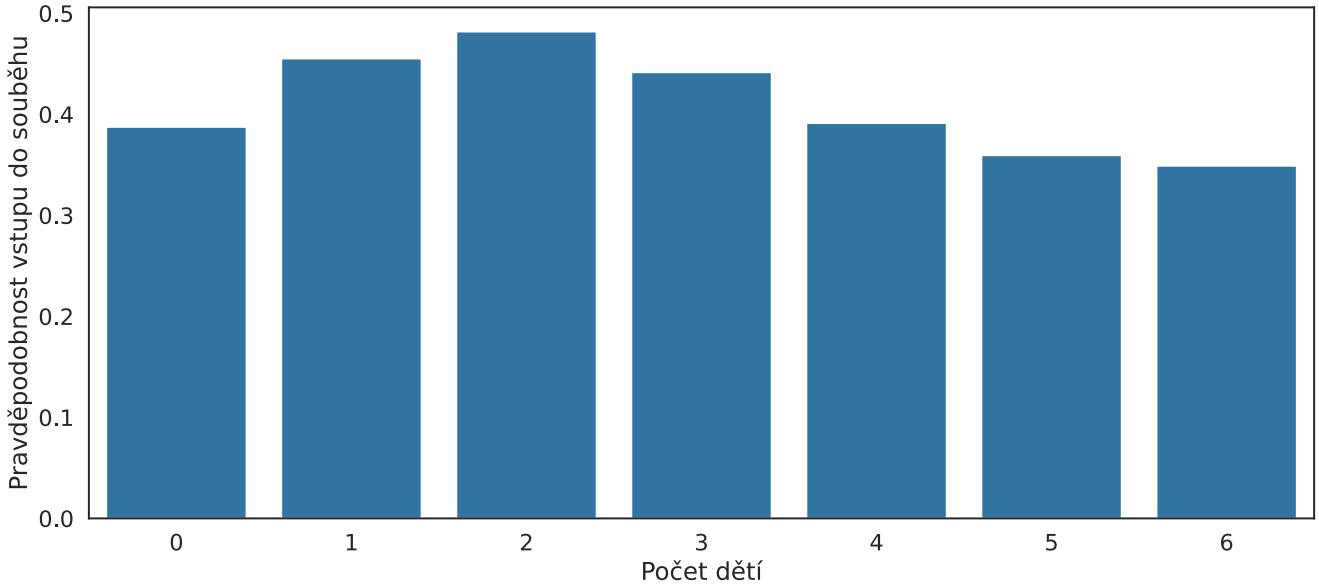

Figure 4.2: Dependence of probability of commencement of concurrence of work with old-age pension at the beginning of retirement on the number of children.

records.

There is no data available based on which marital status could be recognised – i.e., it is not possible to distinguish among single, divorced and married men and women. Information about number of children for men is also not available. In MP database, such information is present, however, according to the documentation, this data was generated stochastically based on occurrence in population and therefore does not carry any value for this analysis.

Factor Impact

Women with 2 children have the highest probability of concurrence of work with retirement, this probability is 48%. The probability decreases for women with both more and with fewer children, see Figure 4.2.

Similar trend can be seen also when analysing transitions from inactivity to employment during retirement which is most frequents also for women with 2 children. The only exception to the rule is

childless women in concurrence – approximately 27% of them transitions to inactivity each year, which

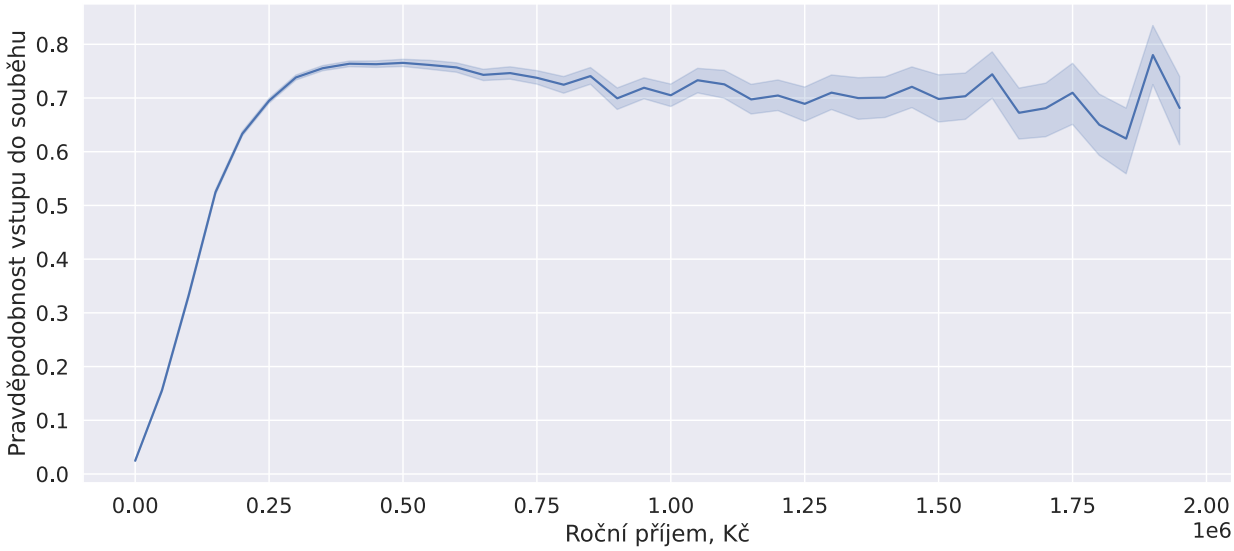

Figure 4.3: Dependence of probability of commencement of concurrence of work with old-age pension at the beginning of retirement on annual income (expressed in million CZK). The curve shows an average value and a 95% confidence interval.

is a higher percentage than for other groups.

Persons receiving widow's or widower's pension commence concurrence less often than those pensioners who do not receive such a type of pension. The difference in probabilities is 5%, specifically 45% and 40% respectively.

The same trend can be observed also in regards to transitions from work to inactivity. While only 0,6% of inactive pensioners who receive both old-age pension and widow's / widower's pension return to work, it is 1,5% for the second group. On the contrary, employment is terminated by 23% of those receiving widow's or widower's pension, while it is 27% for the second group.

Although these factors do not represent ones with the highest impact, taking into consideration good availability of data, we can recommend them for implementation. A disadvantage is unavailability of data needed for impact analysis of number of children for men; i.e., this factor would have to omitted for male part of the population.

4.4.6 Income

Availability of Data

Data about income is immediately available in good quality thanks to STATMIN VZ and STATMIN VZ OSVČ databases.

Factor Impact

Looking at the results of analysis of impact of income on commencement of concurrence, we can see a significant difference between people with low income and people high income. The probability is increasing up to an annual gross income of 300 000 CZK, above this value we observe a slight decrease as you can see in Figure 4.3.

The same conclusion can be drawn from analysis of staying in concurrence of work with retirement, where we can see that people with higher incomes have higher tendency to work.

Based on linear availability of data and significant impact on observed probabilities, we can recommend this factor as the most suitable for implementation.

4.5. Implemented Changes

4.5.1 Preparation of MPs

Entries for implemented factors (education and duration of employment) were already present in model points. Therefore, changes were not necessary. For a more accurate calculation of awarded pension amount, percentage of pension amount established by pension raise (resulting from concurrence) was calculated for already existing model points in retirement.

4.5.2 Prophet

Here, we can divide changes in the model into two parts:

1) Updating tables of transitions between different employment types

Changes were carried out in the following tables: `empl_inact.fac`, `empl_inact_no_event.fac`, `inact_empl_no_event.fac`. These changes included adding the following factors: highest achieved level of education and duration of employment (i.e., duration of time when insurance was paid) in categories 0-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45+ years. Probabilities depend on new factors only in columns associated with commencement of retirement (`PENSIONER`, alternatively `NORMAL_PENSIONER` in table `empl_inact.fac`). There were no changes in other probability columns; i.e., probabilities are constant for various values of highest achieved level of education and of categorised duration of employment.

Newly, return to concurrence from inactivity is enabled, and that is for people who have already reached their retirement age.

2) Pension increase for the reasons of concurrence of work with retirement

A new event `Pension_change` was added into the model. After creating a new random event, it was necessary to increase by one the value of variable `NO_EVENTS`.

Probabilities for event `Pension_change` get applied only at the beginning of the projection (if a given person is a working pensioner at the beginning of the projection), or at the beginning of concurrence of work with retirement in the projection. That means that it is only once determined whether a given working pensioner will be increasing their pension and from there it is assumed that he or she will continue to increase their pension. After applying, the pensioner gets all increases it is entitled to. Pension raise is calculated in variable `PEN_WORK_BONUS_PC`. Probabilities for event `Pension_change` can be found in table `param_cz.fac`.

Into the duration of concurrence of work with retirement, such time is counted when a given person was a pensioner receiving pension and working at the same time. In model points, records are kept of what pension raises resulting from concurrence of work with retirement were awarded to existing pensioners.

4.5.3 Description of .fac Tables

Contrib_period_cat.fac

This table is used for categorizing duration of employment (duration of time when insurance payments were made).

Table 4.1: Structure of Table Contrib_period_cat.fac

Code	Comment
CATEGORY	Duration of employment category
MIN_CONTRIB_PERIOD	Minimum duration of employment for a given category

Empl_inact.fac, empl_inact_no_event.fac, inact_empl_no_event.fac.

These tables are used for determination of probabilities of transitions between types of employment. Added factors are illustrated below; i.e., the illustration does not represent the entire table.

Table 4.2: Structure of Tables Empl_inact.fac, empl_inact_no_event.fac, inact_empl_no_event.fac.

Code	Comment
EDUCATION_MAX	Maximum achieved level of education
CONTRIB_PERIOD_TOTAL	Categorised duration of employment

4.6. Implementation Feasibility Assessment of Other Factors

We estimate the difficulty of implementation of additional factors (type of employment, marital status, income) to be low. It would include primarily changes in tables `empl_inact.fac`, `empl_inact_no_event.fac`, `inact_empl_no_event.fac`, where it would be necessary to add an additional column (or columns) which would include information about the additional factors and determine probability of transitions between different types of employment.

We see a potential problem in regards to longer time needed for calculations – tables are already quite large. Adding additional factors (if they proved significant in the future or if better quality data was available) would result in several times larger table sizes. Therefore, it is important to be wary of the effect on calculation times in case of a potential implementation of additional factors.

4.7. Summary and Evaluation of Other Factors

Table 4.3: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Difficulty of Implementation	Recommendation
Gender	Excellent	High	Implemented	Implemented
Education	Excellent for working persons in extended STATMIN VZ; otherwise poor	High	Implemented	We recommend filling in tables to improve quality of data

Total Duration of Employment	Excellent	High	Implemented	Implemented
Type of Employment	Sufficient for work agreements, employment contracts and self-employment; poor for full-time/part-time	Medium	Low (addition of one column)	We recommend for consideration
Marital Status	Sufficient for widow(er)s and number of children for women; otherwise poor	Medium	Low (addition of one column)	We do not recommend due to insufficient availability of data
Income	Excellent	High	Low (addition of one column, discretization would be necessary)	We recommend for consideration

5. Chapter 2, Commencement of Retirement Pension

5.1. Introduction

In accordance with the results of the feasibility study, an impact analysis of a group of factors affecting commencement of retirement pension and its timing in relation to regular commencement date was carried out. It can be assumed that timing of commencement of retirement pension depends on a situation of a given person. Therefore, the aim here was to extend the current simulation by addition of further factors.

5.2. Data Sources

The following databases were used to analyse commencement of retirement pension and commencement of old-age pension entitlement: INEP database for calculation of insurance duration and STATMIN ANOD database for determination of number of children for women and especially for detection of commencement of retirement pension.

5.3. Impact Analysis of Factors

In the simulation, the event of retirement pension commencement is commanded by table retirement.fac which contains the following:

- Calendar year since when following probabilities will be applied.
- Factors on which the probabilities are dependent, since now only gender.
- Probabilities of commencement of early retirement pension, separately for 1 year to 5 years before the regular commencement date, according to Provision § 31 Act No. 155/1995 Coll.
- Probability of regular commencement of retirement pension (i.e., within one year since commencement of entitlement), according to Provision § 29 Paragraph 1 and Paragraph 3 Letter a) Act No. 155/1995 Coll.
- Probability of postponement of retirement pension commencement for 1 year up to 6 years after commencement of entitlement to retirement pension.
- Probability of late commencement of retirement pension, according to Provision § 29 Paragraph 2 and Paragraph 3 Letter b) Act No. 155/1995 Coll.

In consideration of available data on retirement pensions and absence of major legislative changes, probabilities were entered into the table since year 2010 onwards. Newly, a dependence on career level and a person's replacement ratio was added – this is described in more detail in the next sections. Available data shows that the option of late commencement of retirement pension is rarely used. Furthermore, almost all people who are entitled to late commencement of retirement pension use their entitlement almost immediately and start their old-age pension within the first year. That is why corresponding probabilities remained at value equal to one.

In the following text, we state percentages of people who were entitled to commence retirement pension and also utilised their entitlement. Calculation was performed on monthly basis - number of entitled people divided by total number of people, always taking into account a given combination of factors. Relationship $p_r = 1 - (1 - p_m)^{12}$ was used for conversion from monthly probabilities to annual probabilities used in the simulation.

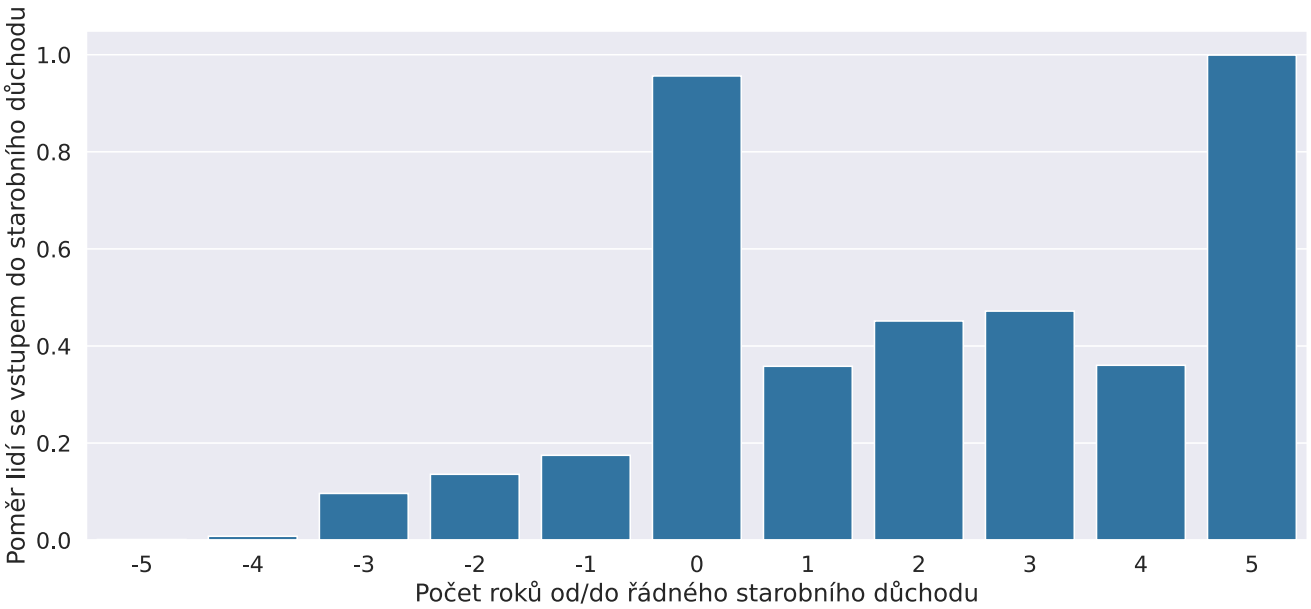

Figure 5.1: Percentage of entitled persons who commenced their old-age pension in a given year (and did not start their retirement in previous years). Old-age pensions with regular commencement date are in point zero on the x axis, early retirements are displayed on the left.

Analysis of retirement pension commencements without incorporating other factors shows that the option of early retirement commencement is used by 36% of those who ever become entitled to it. 14% of those ever entitled to early retirement commencement decide to retire in the year preceding their regular retirement age, and 12% between two and one years prior their regular retirement age. 95,6% of population start their retirement within one year of entitlement commencement. Small number of individuals who have not yet commenced their retirement then start their retirement with probabilities between 35% and 50% in a year. Percentages of entitled persons who commence their old-age pension can be seen in Figure 5.1.

The aforementioned is why in the following description of factor analysis we focus on a percentage of population with early commencement of retirement and also on probabilities on regular commencement date.

Altogether, 9 factors were analysed. Factors of replacement ratio and career level were added to the already implemented factor of gender. Also, an impact analysis of factors was caried out for factors of age, education, marital status, career level of a person's partner, health / sickness rate.

5.3.1 Gender

Availability of Data

Information about gender is present in good quality in each of the databased that were used.

Factor Impact

Probabilities differ a few percentage points between genders. While 37% of men who are newly entitled to early entitlement use this option, for women this percentage is equal to 35%. The percentage of active women who retire in the next year after commencement of their entitlement to regular old-age pension is 95%, same is done by 96% of active men.

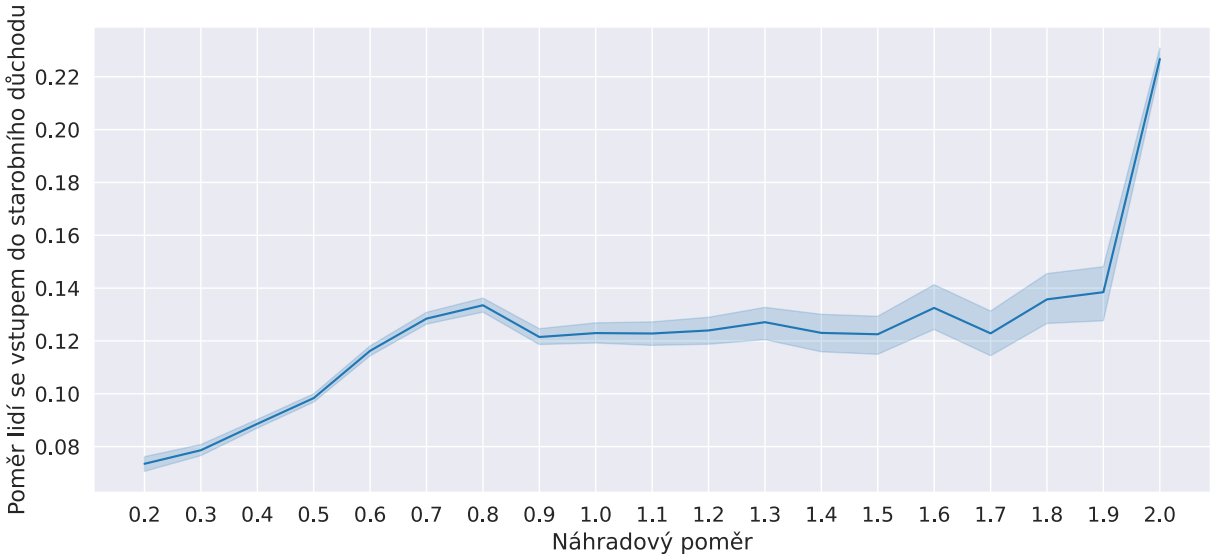

Figure 5.2: Percentage of entitled persons who commenced their old-age pension in the year before their regular commencment date, dependent on their replacement ratio, that is their pension amount divided by their gross salary. The curve shows an average value and

5.3.2 Replacement Ratio

Availability of Data

In the STATMIN VZ and STATMIN VZ OSVČ databases, data about assessment base are easily available. Because of that, records about income are easily accessible. The assessment base for self-employed persons was multiplied by 2, which de facto gives a result of gross income of a person. In case of concurrence of self-employment with another gainful activity, the calculation used a total sum of incomes from the individual gainful activities.

For persons who are actively working, their income was calculated as an average of income for the last 12 months. For people who are not currently working, it was evaluated as unsuitable to set their income at zero - on the contrary, the aim here was to estimate what would be income of such a person if they started to work again. That is why we set *potential income* for them equal to data in the last month of their gainful activity. The same calculation was then used also in Prophet.

The second component of the replacement ration comprises of a pension amount. This is also directly available in ANOD database. The pension amount was not yet known for people whose retirement has not commenced yet. Therefore, this value was calculated according to the valid legislation on percentage assessment of old-age pension in relation to the number of days remaining until regular old-age pension commencement date, or alternatively the number of days after.

Factor Impact

As per expectations, persons with higher replacement ratios are more likely to commence their retirement; i.e., people for whom receiving old-age pension pays off financially. This relationship holds for all relative yearly differences to retirement age. For briefness, only a detail for early commencements of retirement within one year prior to the regular commencement date is mentioned here, see Figure 5.2. This is an overall very important factor and its implementation would have a significant impact on the simulation.

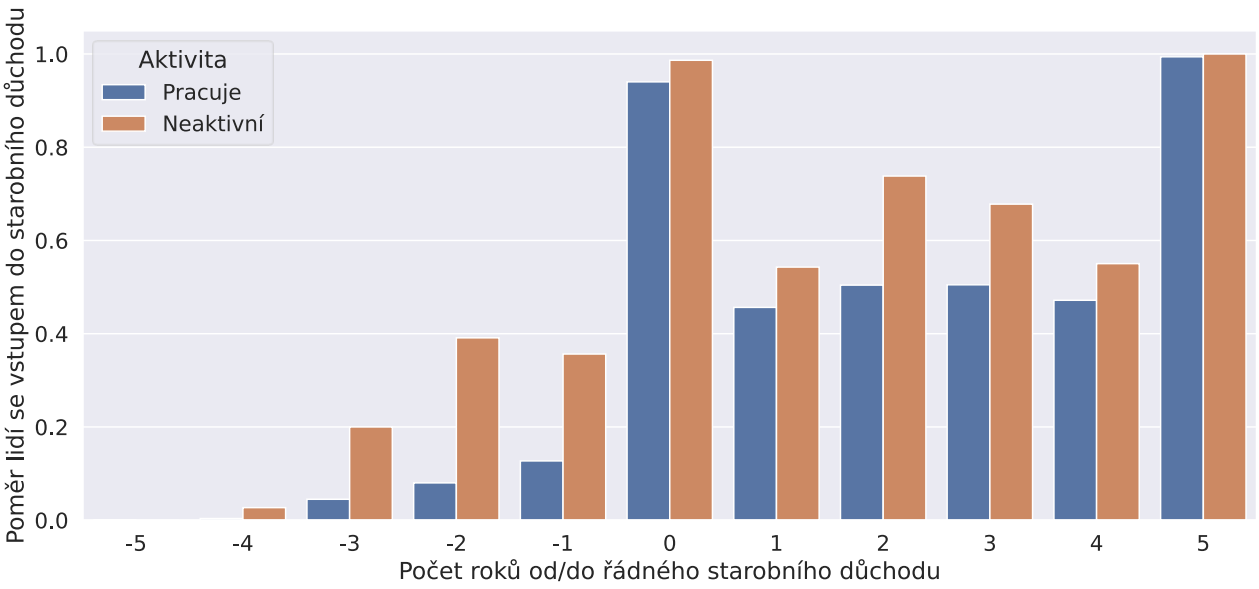

Figure 5.3: Percentage of people who commenced their old-age pension in a given year, dependent on their employment status

5.3.3 Career Level

Availability of Data

Actively working persons can be detected through databases STATMIN VZ and STATMIN VZ OSVČ. In the Prophet simulation, in the remaining part of the population there is a differentiation of inactive persons and unemployed persons, however, we are currently unable to achieve such a differentiation, and that is why we operate only with a binary variable identifying a gainful activity.

Factor Impact

In the illustration Figure 5.3 you can see the impact of this factor on the event of old-age pension commencement, again dependent on the relative number of years to regular retirement age. Herein, the intuitive assumption that people without work commence their retirement more frequently is confirmed yet again. This difference is most noticeable for probabilities of early retirement. For these reasons, we evaluate this factor also as very significant.

Age

The factor of age plays a special role in this context. Given that probabilities of commencement of old-age pension are expressed relative to retirement age, subsequently adding age as a factor loses its meaning. On top of that, the current method of implementation reflects changes in official retirement age, which is why we do not recommend changing it.

5.3.4 Education

Availability of Data

As was already described in Chapter 1, information about education is available only through the extended STATMIN VZ database. However, use of this information is not suitable for the purposes of describing behaviour of the whole population. Given the nature of the available data, it is data only

for the working segment of the population where the commencement of retirement (especially of an early retirement) is less frequent: while 36% of all people entitled to an early retirement use this option, this number is only 25% for an analogical subset drawn only from the extended STATMIN VZ database. When observing regular commencement of retirement, these numbers are more similar – 96% in the entire population and 95% in the extended STATMIN VZ database.

Given the currently available data, this factor cannot be recommended for implementation.

Factor Impact

When we focus only on persons with currently available information about their education, it is apparent that the higher the education, the more probable it is that that person will postpone the commencement of their retirement. While persons with primary or secondary education without "maturita" opt for early retirement at 39% and 34% from those entitled to it, this probability is equal to 20% for persons with "maturita" and lastly, it is only 9% for university educated individuals. In the year of their regular commencement of old-age pension, probabilities of actual commencement is between 95% and 97% for all levels of education except for university education, where the probability is 89%.

Therefore, we can conclude that this factor has a significant impact on the observed events. We recommend it for implementation if better-quality data becomes available.

5.3.5 Marital Status

Availability of Data

In STATMIN ANOD database, data on women's number of children are easily available, and so is information on whether a person is a widow(er). Due to poorer data quality in regard to number of children while taking widow(er)'s pension, the only records which were old-age pension data.

There is no data available based on which marital status could be recognised – i.e., it is not possible to distinguish among single, divorced and married men and women. Information about number of children for men is also not available. In MP database, such information is present, however, according to the documentation, this data was generated stochastically based on occurrence in population and therefore does not carry any value for this analysis.

Factor Impact

The number of children shows an interesting difference between childless women and women with at least one child. 38% of childless women who are entitled to early retirement opt to do so, whereas this number is approximately 33% for women with a child or children. However, as there are very few childless women in the population, this difference should reflect in the old-age pension system without any significant effect. When looking at the next year after regular retirement commencement date, probabilities are almost equal.

People who receive widow(er)'s pension commence early retirement less likely than the rest of the population. Probabilities still remain equal in the third year before retirement age, however, they increase gradually from here and one year before retirement age old-age pension is commenced by 12% of those who receive widow(er)'s pension, while it is 18% for the rest of the population. Again, as widow(er)'s pension is received only by a small number of individuals, the difference herein only plays a small role.

It would be necessary to obtain further data in order to carry out a complete analysis of this factor. Taking into account that the impact of currently available variables is rather small, this factor is considered to be one of the less interesting factors for implementation.

5.3.6 Career Level of a Person's Partner

Availability of Data

This factor is not available in the data at all. Records about a person's partner can be found in MP database, however, according to the documentation, this data was generated randomly and it is therefore unusable for a determinative analysis.

Factor Impact

Given the unavailability of data, the analysis could not be carried out.

5.3.7 Occupation

Availability of Data

Information about occupation is available through the extended STATMIN VZ database.

Factor Impact

The analysis showed significant differences in behaviour of people with different occupations. A potential detailed analysis or implementation cannot be carried out due to the fact that probabilities of early retirement for those persons for whom the information about occupation is available are lower than probabilities for the entire working population.

While according to the STATMIN VZ database, 18% of working people entitled to early retirement retire early, for persons from the extended STATMIN VZ database it is 8%. Taking into consideration the fact that contracts in the extended STATMIN VZ database constitute a subset of contracts in the STATMIN VZ database, we must state that the extended STATMIN VZ database does not include a random sample of population, and that is why it cannot be used for an analysis where results should describe the entire population. Therefore, this factor with the currently available data cannot be recommended for implementation.

Nevertheless, let's look at the available results. When looking at the first level of occupational codebook CZ ISCO, we can see a large difference among categories in regard to the frequency of early retirement commencement: this option is most used by ancillary and unskilled workers (14%); followed by workers in agriculture, forestry and fishery (11%); workers in services and sales (11 %) and people operating machinery (10%). On the contrary, early retirement is very rare for lawmakers and managerial roles (2%), specialists (3%) and specialised workers (4%).

Similar trend holds also for regular commencement of retirement, which is most frequent for workers operating machinery and craftsmen (both 97%) and the least frequent again for lawmakers (90%).

5.3.8 Health / Sickness Rate

Availability of Data

Data on sickness rates are available in the SEE20 database and its connection to the INEP database through the DELNEZ connector. The SEE20 database contains instances of sick leave mostly granted

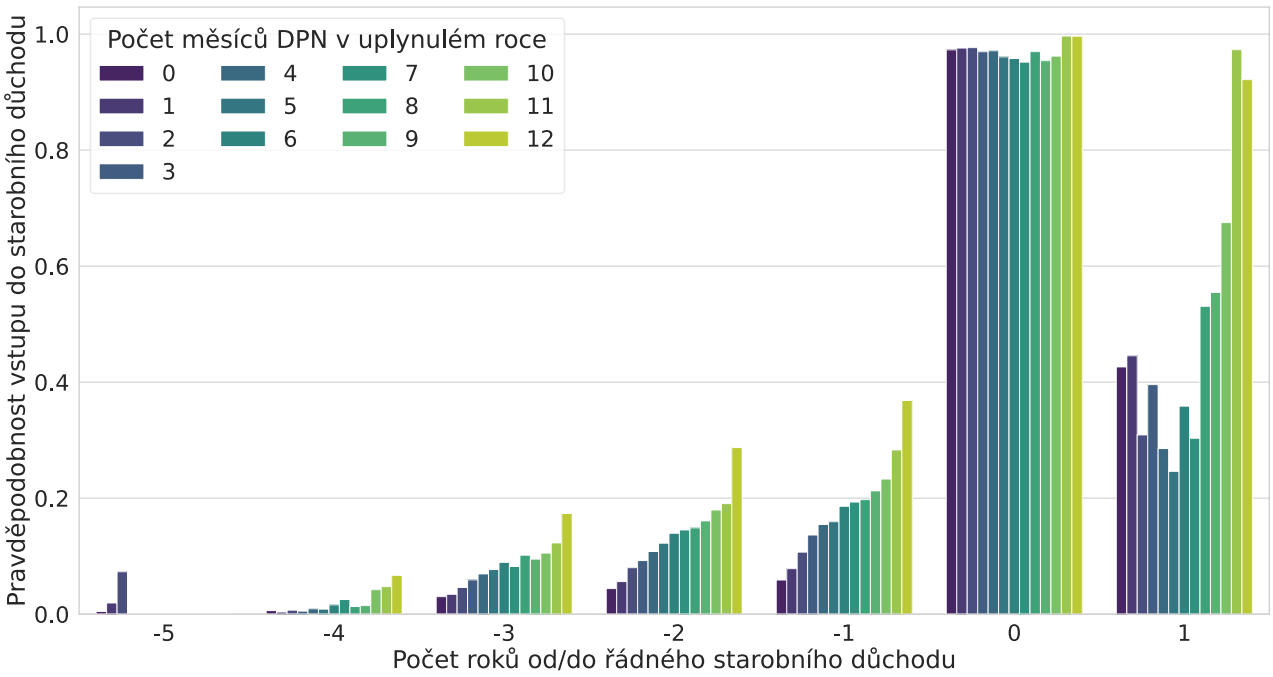

Figure 5.4: Percentage of people commencing retirement in the given year dependent on the total number of months spent on sickness leave during the last year.

to employees. To be able to describe complete medical history for every person, only employees were analysed.

Factor Impact

Figure 5.4 shows the dependence of commencing retirement depended on the number of months spent on sickness leave in the last year. As per expectations, people with longer periods of incapacity for work commence their retirement sooner. Especially high are probabilities of commencing retirement after a 12 months long sickness leave.

This factor is one of the most significant out of those considered.

5.4. Implemented Changes

5.4.1 Preparation of MPs

All necessary information was already available in the model points that enter the simulation.

Prophet

Here, only a change of table `retirement.fac` was carried out. Two new factors determining probability of old-age pension commencement were added into the table – the replacement ratio and an indicator of employment (indicating whether a person is employed or not).

The replacement ratio for an employed person (including self-employment) is calculated as the forecasted full monthly old-age pension divided by average of gross monthly income earned in the last 12 months. The replacement ratio for unemployed and inactive persons is calculated as the estimated full monthly old-age pension divided by the last earned monthly income.

The second factor determined only whether a person is employed or not (unemployed or inactive person). Thus, column `EMPLOYED` can have value 1 (employed) or 0 (unemployed / inactive).

Moreover, variables loading up values from tables were edited and a table enabling categorisation of the replacement ratio (as described above) was added.

5.4.2 Description of .fac Tables

Retirement.fac

This table is used for determination of commencement of retirement probabilities. Added factors are displayed below, the entire table is not illustrated.

Table 5.1: Structre of Table Retirement.fac

Code	Comment
PEN_GRS_SAL_RATIO_CAT	Categorised replacement ratio
EMPLOYED	Indicator determining whether a given person is employed

Retirement_pen_grs_sal_ratio.fac

This table is used for categorisation of the replacement ratio.

Table 5.2: Structure of Table Retirement_pen_grs_sal_ratio.fac

Code	Comment
CATEGORY	Replacement ratio category
MIN_PEN_GRS_SAL_RATIO	Minimum replacement ratio for a given category (one decimal place at most)

5.5. Implementation Feasibility Assessment of Other Factors

The implementation difficulty of additional factors (listed in table below) should be low. Primarily, it would consist of changes in table `retirement.fac`, where an additional column (or columns) with information about the additional factors and determining probabilities of old-age pension commencement would have to be added.

We see a potential problem in regard to longer time needed for calculations – the table is already quite large. Adding additional factors (if they proved significant in the future or if better quality data was available) would result in several times larger table sizes. Therefore, it is important to be wary of the effect on calculation times in case of a potential implementation of additional factors.

5.6. Summary and Evaluation of Other Factors

Let's compare the above-mentioned results with the outcomes of project "Identification of socio-demographic characteristics influencing timing of retirement commencement" (Šlapák, Holub, Průša, 2017), within which answers to a sample survey in the form of a questionnaire which was filled in by 805 805 respondents together with their application for old-age pension. The aim was to map the influence of factors significant in retirement commencement decision-making, some of these factors are identical with those described above.

Results of an analysis of the survey correspond in many aspects with the results of the herein described data analysis. People with "maturita" or university education tend to opt for early retirement

substantially less often than those with lower education. Same holds when comparing results for the factor of career level, also here we obtain identical results; i.e., higher probability of early retirement for unemployed persons. Similarly, with occupation and income (in this document used for the replacement ratio) we can see corresponding postponements of retirement commencement for specialised roles and roles with better financial evaluation.

Contrary to the above-described analysis, the survey did not prove a dependency of retirement commencement on a person's health condition.

In comparison with the herein analysed factors, the survey included information about marital status, which showed as an important factor for timing of retirement commencement as single people tend to retire more often.

Moreover, an analysis of differences between employees and self-employed was carried out as a part of the project, also postponements of retirement among self-employed persons was described.

Table 5.3: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Gender	Excellent	High	Implemented	Implemented
Replacement Ratio	Good (non--trivial transformation of available data)	High	Implemented	Implemented
Career Level	Good (differentiation between unemployed and inactive persons is not possible)	High	Implemented	Implemented
Age	Excellent	-	High (rewriting of the entire chapter)	Keep the dependency on from/to retirement age
Education	Excellent for working persons; otherwise poor	High	Low	We recommend for consideration if better-quality data is available
Marital Status	Sufficient for widow(er)s and number of children for women; otherwise poor	Low	Low	Given incomplete data and low impact, we do not recommend this factor for implementation
Career Level of a Person's Partner	Data does not exist	-	-	Cannot be recommended
Occupation	Poor, data from extended STATMIN VZ do not include a random sample of the population	High	Low	Cannot be recommended with currently available data
Health / Sickness Rate	Good for employees (connection through DELNEZ necessary)	High	Low	We recommend the factor for implementation

6. Chapter 3, Care for Dependents

6.1. Introduction

The aim of this chapter was to implement the process of care for dependents in degrees of dependence I – IV and use it to replace the current structure and decision process (commencement and termination) of care for family.

6.2. Data Sources

The fundamental (and the only) source of data for this chapter was database PnP (abbreviated "Příspěvek na péči", in English care allowance). This database includes information about individual cases where benefits were rewarded. Each row includes especially an ID of the caregiver and an ID of the dependent, year and month of birth of both persons, gender and address of the caregiver, degree of dependence, date of the start and closure of the case and, if applicable, also the date of passing of the dependent.

Database PnP covers period from the beginning of year 2007 until August of year 2020. This timeframe was shortened to 2014 - 2019 for some analysis. There was a change in supplier of the database system in year 2014 which disrupted the consistency of data to an extent, which would give distorted results to some analyses. The fact that data for 2020 were not available for the entire year causes problems especially with calculations of probabilities of various events.

Identification details available in this database (ID of persons) do not match the data in other databases which were used (such as database INEP, STATMIN VZ, etc.) – therefore database PnP could not have been connected to these data sources. Consequently, no other database could have been used.

Potentially joining databases PnP and INEP by matching identifiers would likely significantly improve the understanding of economic behaviour of individuals who care for dependents. It would be for example possible to analyse the impact of the degree of dependence (and its changes) on gainful activity of the caregiver, probability of early retirement, change in occupation and other important events. It would be also possible to significantly improve the method of model point allocations. Currently, the status "caregiver" is assigned stochastically in such manner that gives a corresponding total number of caregivers in the population. If identifier matching was carried out, it would be possible to assign the caregiver status to real people in the model point database and by doing so the default state of the simulation in regard to this chapter would be significantly better.

6.2.1 Preparation and database cleansing

Removal of Duplicate Benefit Cases

Due to changes in suppliers of the database systems and thus changes in the benefits database systems, multiple records for the same event were created in the database on many occasions. As a result, one benefit case can have multiple reference numbers. Similar situation occurs at the event that a caregiver moves house – this can result in transfer of the relevant file in the new location and it there being assigned a new reference number. A necessary step preceding analyses was to delete such multiplications.

A combination of columns `PEC_ID`, `ZDT_ID`, `OD`, `DO` is used for unambiguous identification of a benefit case. The basic principle of database cleansing is that for each pair of rows which does not differ in

the mentioned columns but differs in reference numbers, we identify duplicate reference numbers. From these gathered pairs of reference numbers, we then decide which one is "worse" – i.e., which one is older (the middle part of the reference number can be interpreted as the year when the file was created) or which is less complete (present in fewer records), and we delete all rows where this less suitable reference number is used.

By following the above-described process, we obtain a database which contains approximately 90% of the original records (1.85 million records in comparison to originally 2.06 million records). This cleansed database does not contain any two records which would be identical in all the above-mentioned identification columns. While performing the process of database cleansing, no dependent was lost, however, approximately four thousand caregivers were lost in the process (total number of unique caregivers is approximately 1 million, the loss amounts to 0.4% from this number). Based on several manually verified records, we assume that the "lost" caregivers are usually caregivers who shared care for the dependent simultaneously with another caregiver under the same reference number and that such concurrence of care occurred in the past and was not transferred to the new database system in its full content.

Preparation of Tables with Monthly Granularity

Most analyses rely on data with monthly granularity. The original database PnP, on the other hand, includes two columns `OD`, `DO` (alternatively `ZDT_DATUM_UMRTI`) which state when the benefit case was created and when it was closed. It holds that the duration of a benefit case is not strictly limited and can continue for even many years. The most common reasons for closing a benefit case are the termination of care, change in the degree of dependence or for a similar reason, unrelated to duration. Because it generally holds that one person can subsequently or simultaneously take care of more dependents and one dependent can have subsequently or simultaneously more caregivers, it is necessary to prepare two different tables with monthly granularity for the analyses – one from the viewpoint of the pair caregiver-dependent and another from the perspective of dependents.

The table from caregiver-dependent perspective includes three unique columns – ID of the caregiver, ID the dependent and a month in which care was given.

Table from the perspective of dependents includes only two unique columns – ID of the dependent and a month in which care was given. A problem which had to be solved here was that some ID-month combinations stated multiple degrees of dependence according to different benefit cases in the original database PnP. This occurred for approximately 6% of ID-month combinations. Logically, at any given time each dependent can be in just one degree of dependence, which is most likely the degree which was assigned to the more recently given benefit case preceding the month when a person appears to be in multiple degrees of dependence. For example, a person in year 2019 is appearing to be simultaneously receiving benefits for third degree of dependence and for fourth degree of dependence – third degree of dependence benefits started in year 2011 and fourth degree of dependence benefits started in year 2016 – it is most likely that the true health condition is reflected by the most recently awarded benefits, in this case dependence degree 4.

Monthly granularity of the database naturally leads to monthly probabilities of transition. However, event probabilities in Prophet are yearly probabilities. The conversation is carried out by the following power relationship: $p_r = 1 - (1 - p_m)^{12}$.

6.3. Collective Care

In order to suggest the implementation of care for dependents into model NEMO, it was necessary to analyse how frequent is a situation which does not correspond to the simplest model scenario of one caregiver – one dependent in one unbroken period. There are multiple other possible scenarios when it comes to care for dependents.

6.3.1 Interrupted Care

Interrupted care describes a situation when a caregiver terminates care for a dependent but after some time they again start to care for the same person, in such a case we disregard whether the dependent received care by another person during the interim period. The situation can also be analysed in reverse, i.e., the dependent ceases their dependence status but after some time, care from the same caregiver is resumed.

Interruption of care by the caregiver is a relatively marginally occurring event that only occurred for approximately 1.5 % of all caregivers. Approximately a half of those then resume their care within one year.

Interruption of care by the dependent is a similarly rare occurrence, approximately 2 % of dependents interrupted the care, and approximately two thirds of those then resumed the care within a year since the interruption.

Interruption of care by both the dependent and the caregiver is not explicitly modelled in the model, each benefit case is considered individually. Due to the limited number of people for whom a given caregiver can care, it is likely that factual interruptions of care will occur, i.e., a reinstated caregiver status does not determine that the newly commenced care is specifically interrupted care, but it indicates it by logic.

6.3.2 Concurrent and Subsequent Care

Concurrent care describes a situation when one caregiver gives care for more than one dependent simultaneously. Subsequent care describes a situation when one caregiver takes care of more dependents but subsequently (one after the other) not simultaneously.

This occurrence is also relatively rare, approximately 88 % of caregivers have never cared for more than one dependent. The number of caregivers who have cared for more than two dependents is marginal (approximately 1 % of caregivers). The remaining 11 % of caregivers has experience with taking care of two dependents. Approximately half of them provided care exclusively subsequently (i.e., they never concurrently cared for two persons). Only 1 % of all caregivers cared for two dependents during the entire duration of their caregiver status.

For both concurrent and subsequent care, the most common model of care is care for two persons both of whom are a generation older (parents). A common model of a subsequent care is also caring for one person who is a generation older and for a second person of the same generation (i.e., partner of the caregiver). In case of concurrent care, it is also quite common to be a caregiver of two children (i.e., caring for persons who are a generation younger).

Due to relative marginality of the concurrent care, concurrent care is not simulated in the model. On the other hand, the model of subsequent care is supported in the model – a caregiver who has terminated caring for one dependent can start caring for another person.

6.3.3 Shared Care

The term shared care is used to describe a situation when multiple caregivers (subsequently or simultaneously) care for one dependent. This situation is more common than the previously-described type of care – approximately a quarter of all dependents have experienced this type of care. A typical example of shared care is a situation when one older dependent (usually a parent or a partner) is being cared for by multiple caregivers for the entire duration of the care (unbroken time period). On the other hand, the most common model of care for children is a subsequent care when one caregiver replaces the other caregiver (in most cases parents of the child) at some point in time.

This type of care is not enabled in model NEMO for principled reasons (dependents are modelled as secondary and independently for each main person). The decision to omit this care type from the model can potentially lead to higher number of dependents during simulation, however, this fact does not influence further calculations in the current model (care allowance is not modelled).

6.3.4 Summary of Collective Care

The implementation of dependents (see below) is similar to the implementation process of children – each individual whose life path is simulated (the main person) is assigned multiple secondary persons and from those, a dependent is selected. This approach basically makes some types of collective care impossible (especially shared care), while other types (such as subsequent care) are fully supported.

Interrupted care (i.e., termination of care and then resuming care in the same caregiver-dependent combination after some time) is possible, however, the duration of the interim period is not modelled. Reasons for interrupted care are for example if an individual cares for their partner, the care is terminated for other reasons than for decease, after some time giving care is resumed and the caregiver's partner is again selected as the dependent. It is practically insignificant that the dependent is the exact same person rather than the acknowledgement that the dependent is a person of the same age and gender.

Concurrent care (i.e., the caregiver cares for two or more persons simultaneously) is not currently supported in the model – each person can be assigned only one dependent. In regard to the pension system, it is insignificant whether the individual cares for one dependent or for more dependents – it is the dependent who gets awarded care allowance, a potential payment towards health insurance and inclusion of insurance replacement period do not depend on the number of dependents a caregiver cares for. Theoretically, the simulation of concurrent care could be implemented, however, we do not recommend it due to the above stated reasons.

A subsequent care (i.e., a caregiver cares for two or more persons subsequently, not simultaneously) represents an analogical situation to the interrupted care, in this case, however, the dependents are not the same person. This type of collective care is enabled in the model, but it is not explicitly modelled. At the moment when a caregiver terminates a benefit case and stops providing care to the dependent, the caregiver again becomes an "ordinary" person who can start providing care for a dependent. When making a new decision whether this person will become a caregiver, the fact that they have been a caregiver in the past does not play any role.

Shared care (i.e., one dependent is being cared for by two or mode caregivers) is not supported in the model and it also cannot be supported since main persons do not interact with each other, which means that it is not possible to ensure that two benefit cases would be mutually synchronised.

Collective care is a situation which occurs in reality, but with respect to the pension system, it is not very significant. A much more significant event is the fact that a person is a caregiver at all. For the model, an interesting trait of the dependent is their age and gender as these factors determine a potential transition to a different degree of dependence and also the termination of care.

6.4. Factor Impact Analysis

Within the chapter 'Care for dependents', factors of age (for both the caregiver and the dependent), gender of the caregiver and a region of the caregiver were analysed. Factors which were implemented were the factor of age and gender.

A caregiver status when a caregiver cares for a person close to them is in model NEMO operated by a set of tables with probabilities of various events which are described below. These probabilities depend on a combination of various factors. The most significant factors are the age of a caregiver, the age of the dependent and the gender of the caregiver. The most significant events here are the beginning of care, the termination of care and also the transition from one degree of dependence to another. The analysis also focused on the method of assigning dependents to caregivers.

6.4.1 Age

Data Availability

The age factor is in database PnP available in high quality, both information on the year of birth and information on the month of birth are available for both the caregiver and the dependent.

Factor Impact

The age factor is a crucial factor for the entire chapter, and it enters all events. Besides the age of the caregiver and the age of the dependent, it is also interesting to focus on the difference between their ages, i.e., the relative age of the dependent.

Age of the caregiver is used as a factor only for the event of care commencement (see Figure 6.1). This probability gets evaluated in the simulation every month for all individual who are not caregivers and are 15 years of age and older. The highest probability of care commencement is for women in pre-pension age, probably in relation to caring for older parents.

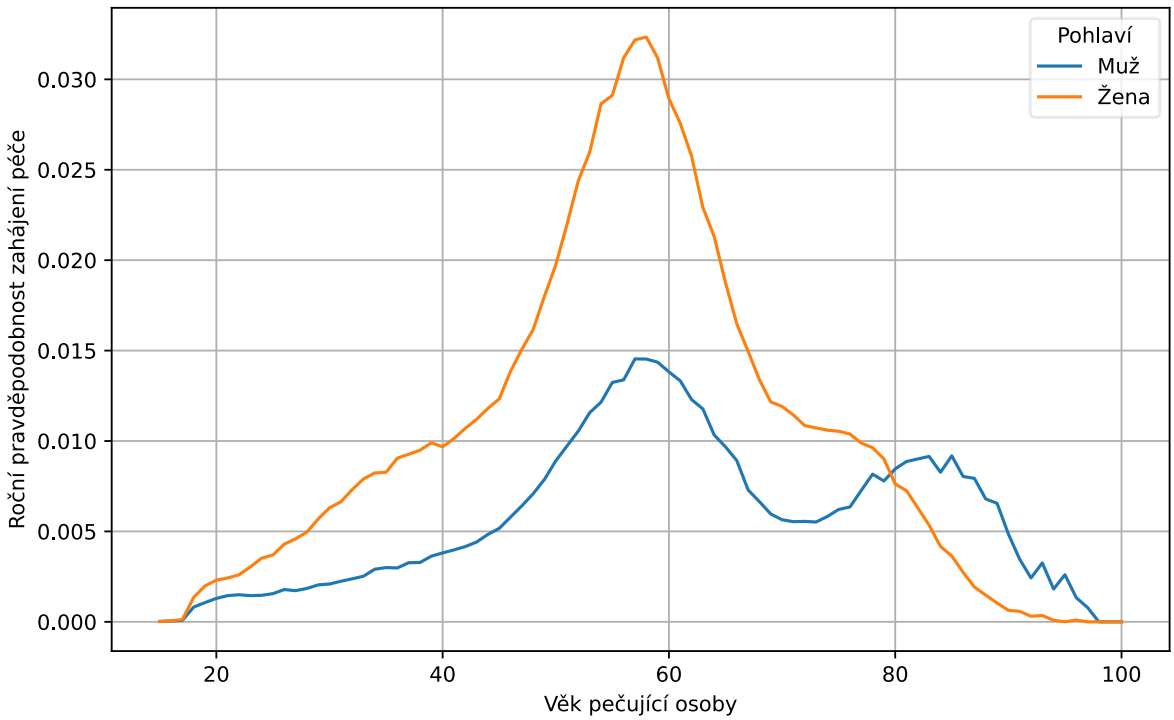

Figure 6.1: Probability of care commencement in dependence on age and gender of the caregiver.

Age of the dependent is used for evaluations of probabilities of care commencement, care termination (due to passing as well as for other reasons) and for changes in the degree of dependence. The probability of care commencement depends on both the age of the dependent and the degree of dependence which they commence. This probability is used for selection of a dependent from a set of admissible persons (see Implemented Changes).

The event of change of the degree of dependence (Figure 6.2) is evaluated every month for each dependent. In this case, the factor of age is divided into three categories (1-19, 20-64, 65+) in such a way that gives sufficient data in each category while each category covers a group of people with consistent behaviour (children / youth, adults, pensioners). In the youngest age category, it holds that is relatively common to transition into lower degrees of dependence while in the oldest category it is much more common that the health condition worsens which results in a transition into a higher degree of dependence.

Termination of care (Figure 6.3) is possible in two fundamentally different ways – due to passing of the dependent, or for a reason other than passing of the dependent. The probability of passing of a healthy individual (for example a standard main person) is based on mortality tables which describe the probability of decease dependent on age and gender of a person. For dependents, it is necessary to amend these probabilities by a coeficcient which describes “excess mortality”, i.e., how many times higher the probability of decease is for a person in the given degree of dependence compared to a healthy person of the same age. The excess-mortality coeficcient is calculated as a ratio of probability of passing of the dependent (derived from database PnP based on column `ZDT_DATUM_UMRTI)` and probability of passing of a healthy person of the same age, same gender and same year of birth (mortality tables `Mort_females_CSU_2018.fac` and `Mort_males_CSU_2018.fac`).

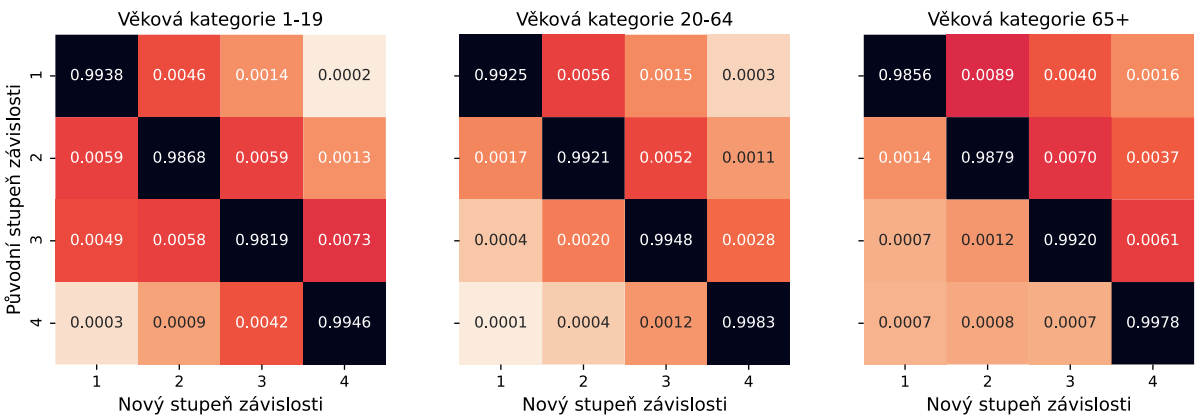

Figure 6.2: Monthly probability of change in the degree of dependence in relation to age of the dependent and to a previous degree of dependence.

Termination of care for a reason other than passing depends on the age of the dependent and the degree of dependence. All degrees of dependence (with the exception of degree 1) have increasing probability of termination of care with increasing age and peak around the age of 80. On the other hand, a peak for degree 1 is around 18 years of age when legislative conditions for receiving care change and approximately a sixth of all persons in the first degree of dependence stop being dependent at this age.

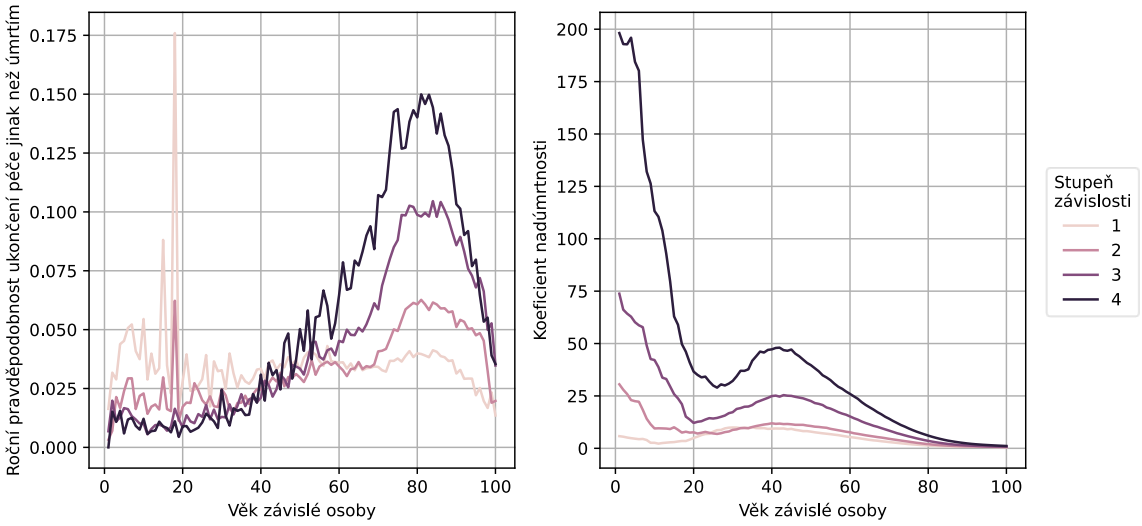

Figure 6.3: An analysis of termination of care in relation to age of the dependent. Yearly probability of termination of care for a reason other than passing of the dependent (left panel), excess-mortality coefficient for dependents (right panel).

Interesting to observe is an analysis of relative age of the dependent (Figure 6.3). The graph shows three significant peaks (and one smaller peak) – these peaks correspond to generation differences between caregivers and dependents. Boundaries between individual generations are approximately 15 years apart. The left peak (when a dependent is more than 15 years younger than their caregiver) corresponds typically to the situation when a parent takes care of their dependent child, this is approximately 10% of all benefit cases. The middle peak (approximately 30% of cases) corresponds to care for a partner or a sibling. The highest peak in the graph (approximately 50% of cases) typically

corresponds to children caring for older parents. The smaller peak, located on the right-hand side of the above-described significant peaks, is around 50 years of difference in age, which typically corresponds to grandchildren caring for their grandparents, this is less than 10% of cases. The distribution of differences in ages corresponding to children caring for their parents and grandchildren caring for their grandparents was used when generating parents as potential dependents.

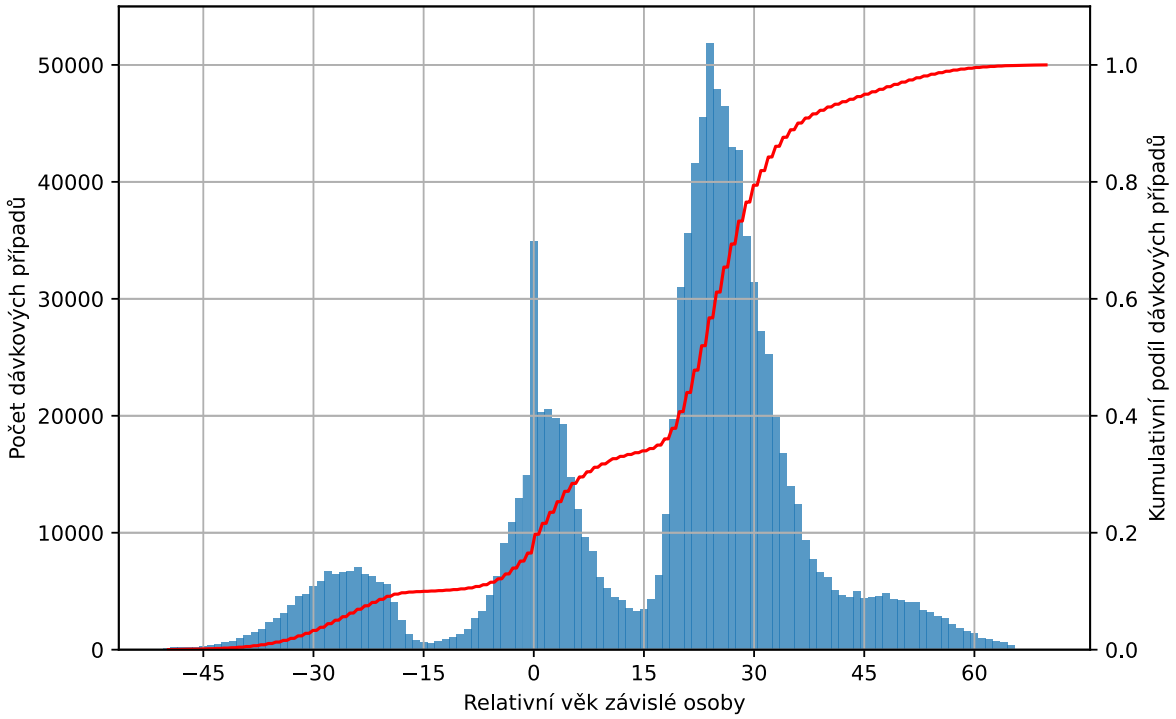

Figure 6.4: A histogram of relative age of a dependent (blue area, left axis) and a cumulative ratio from the total number of cases (red line, right axis).

6.4.2 Gender

Data Availability

Database PnP includes only information about the gender of a caregiver (in high quality), does not include information about the gender of a dependent.

Factor Impact

The gender of a caregiver enters the model as a factor only at time of care commencement, other processes and statuses are evaluated from the perspective of a dependent for whom the gender is unknown.

The probability of care commencement (see Figure 6.1) significantly differs for each gender – the probability of care commencement for women is most of the time higher than for men, with a significant peak around 60 years of age. The curve for men is lower, men comprise only approximately 30 % of all caregivers, but it has two distinct peaks. Apart from increased probability around 60 years of age, there is also a distinct peak around the age of 80 for men, this second peak is not observed for women.

We assume a significant impact of the factor of the dependent's gender, especially when deciding to care for older parents which is when the demographic curve differs significantly for women and men.

6.4.3 Region

Data Availability

Information about region is available in database PnP in the form of a postal code (PSČ), municipality and district of the caregiver. All these columns are complete for all rows in the database, i.e., the availability of data is excellent.

Factor Impact

The factor of region was analysed from the perspective of a ratio of caregivers in the population. This approach avoids dependence on gender and age of the caregiver and of the dependent. We assume that these dependencies do not differ much across regions and thus, including them in the analysis would significantly complicate interpretation.

Regional information from database STATMIN VZ was used as a basis population for the analysis of location because correctly assigning location requires population at the level of individual postal codes (PSČ), these are not otherwise available. This approach inevitably leads to an overvaluation of the ratio of caregivers because database STATMIN VZ contains only a part of the population (employees), however, such overvaluation occurs relatively evenly for all regions. For analysis of districts, a table of population from the Czech Statistical Office (ČSÚ) (ČSÚ, 2021) was used, this table can bring in a small inaccuracy for districts where the number of inhabitants significantly fluctuated in the evaluated years (2014-2019).

Comparing the ratio of caregivers according to locality (Figure 6.5) shows that the distribution of caregivers is practically the same for rural and for urban areas (F-test gives $p > 0.05$ for each degree of dependence).

No interesting patterns were observed at the level of districts (Figure 6.6). The only practically interesting element on the map is the district of Rakovník which has an above-average ratio of caregivers in most degrees of dependence. On the other hand, the rest of the Central Bohemian Region shows rather below-average ratios. Other districts are not mutually compared, alternatively they show out-of-average ratios only in one degree of dependence (for example, Kroměříž only in degree 1) where it is likely that the deviations are only random fluctuations.

The impact of both subfactors (locality and district) is negligible and thus, we do not recommend the factor of region for implementation.

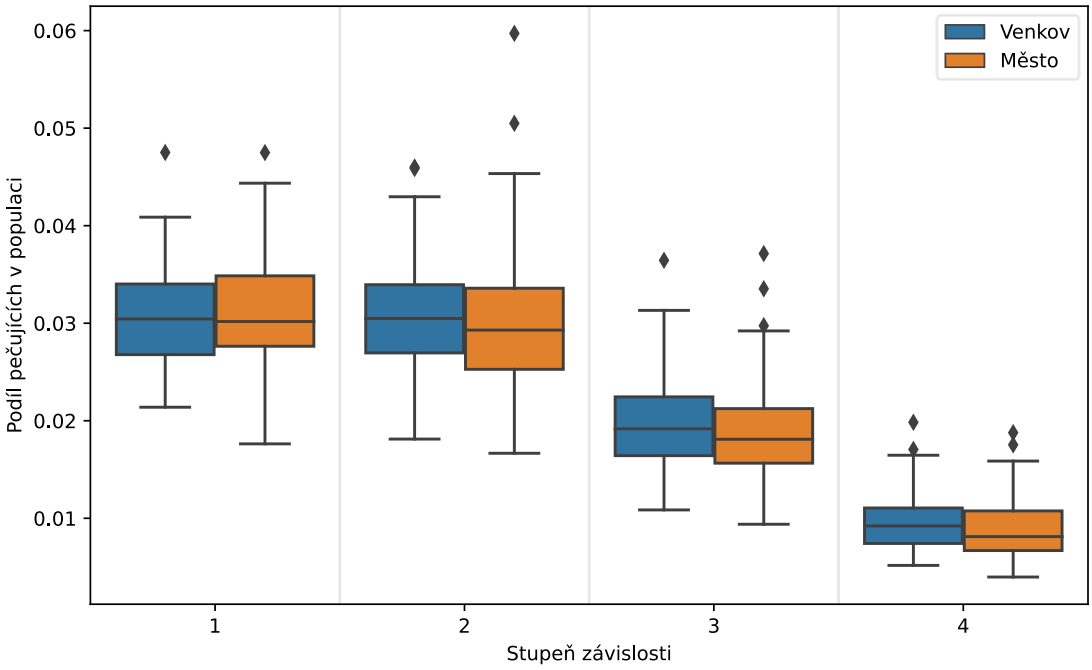

Figure 6.5: Ratio of caregivers in population in rural and urban areas in relation to the degree of dependence of the dependent. The difference among localities is insignificant for all degrees of dependence.

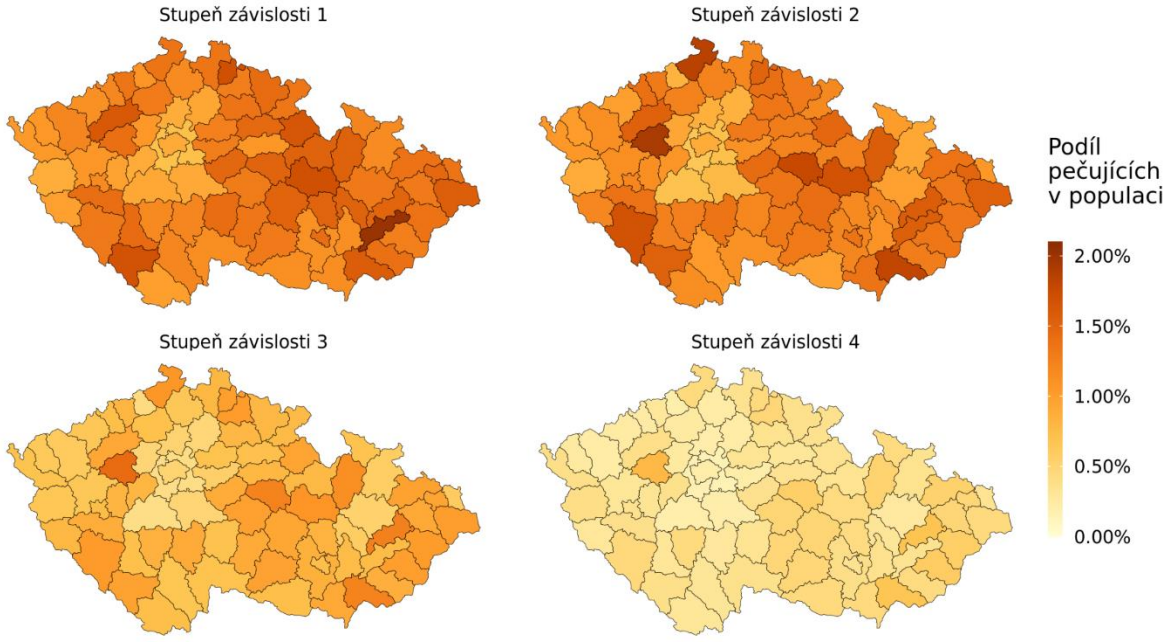

Figure 6.6: Ratio of caregivers in population of individual districts. Each map corresponds to one degree of dependence.

6.5. Implemented Changes

6.5.1 MP Preparation

No new columns were added to the model point database because it is not possible to unambiguously determine who is a caregiver (or a dependent) due to incompatibility of IDs between databases MP and PnP. Therefore, the implementation of generating dependants based on probabilities and the number of dependents in DCS was performed.

The process of generating dependents is divided into two phases – a main phase and a corrective phase. The main phase is implemented in DCS immediately after pairing with a partner (`08_partners_pairing.DCS`). The corrective phase is implemented in DCS `09_SPCODES.DCS` due to the necessity of repeated passing though all model points.

The first step, before a dependent is determined, is to assign both parents to each main person in MP. This step is implemented in `08_partners_pairing.DCS`. Each main person is assigned three new fields containing information about parents:

- `INIT_PARENT_AGE[2]` – age of both parents
- `INIT_PARENT_SEX[2]` – gender of both parents
- `INIT_PARENT_STATUS[2]` – identifier whether the parents are alive (`INIT_PARENT_STATUS = 0`)

The process of generating parents for all main persons is based on the probability of age difference between the main person and each parent, with the use of a newly added FAC table `prob_parent_age_diff.fac`. The obtained age difference for each parent is then added to the age of the main person, this determines the age of both parents. In order to determine whether the parents are still alive, probability in table `life_expect_rates.fac` is used.

After pairing with parents, the determination whether the main person will be a caregiver is then carried out in the code. Only living main persons in the age from 15 to 100 years can care for a dependent. If a person is assigned to be providing care, it is necessary to find out which kind of care they provide (care for a child / for parents / for a partner). This decision is driven by probabilities in table `prob_person_care.fac`. The table contains probabilities that a caregiver will provide a certain type of care based on their age and gender. Before further assessment of the type of care, it is verified whether a potential caregiver has dependents at the given time, for whom it is possible to care (for example, in case of caring for a child, it is necessary to verify, whether the caregiver has children and whether at least one of their children is alive).

As soon as a type of care is allocated, the caregiver is paired with a dependent and a degree of dependence is selected. Table `person_care_count_age_cat.fac` contains real numbers of dependents, divided into three age groups (0 – 18 years, 19 – 65 years and 66+ years) according to gender and an assigned degree of dependence. The dependent is selected differently for each type of care:

- Care for a child – for each living child, the total number of possible dependents in the given age category according to gender and for each degree of dependence. Based on these numbers, a probability of care is calculated for each considered child in the given degree of dependence. The child with the highest probability in the given degree of dependence is selected as a dependent.

- Care for parents – similarly as it is for children, each living parent, based on the number of available dependents, gets calculated a probability that they will be a caregiver who will take care of this parent in the given degree of dependence. The parent with the highest probability in the given degree of dependence is selected as a dependent.
- Care for a partner – only a degree of dependence is chosen for partners – again based on the number of dependents in the given degree of dependence and on an age category. Because a care for a partner is not conditioned by marriage, the couples considered are besides married couples also couples who have a child or children.

After a successful selection of the correct dependent in the given degree of dependence, the number of available dependents in table `person_care_count_age_cat.fac` is reduced by one. If, during the selection, a situation occurs when the selected dependent (child / partner / parent) has already used up the total number of possible assigned persons in all degrees of dependence, all attributes that the main person cares for someone are removed – despite the fact that the probability states that the main person should be caring for someone. Caregivers who were assigned care (table `caregivers_count.fac`) get deducted similarly. To monitor the number of assigned dependents, three age groups with different behaviours were chosen – the reason for age groups was that when an accurate age was used, many cases lacked potential dependents and so very few caregivers were generated.

The result of a successful selection of a dependent is comprised of two new variables which are added into the output MP:

- `INIT_PER_CARE_DEPENDANT` – an identification of a dependent
  - 0 → no dependent (i.e., the main person does not care for any person)
  - 1 → care for a partner
  - 2 → care for a first parent
  - 3 → care for a second parent
  - 4 → care for a first child
  - …
  - 8 → care for a fifth child
- `INIT_PER_CARE_LEVEL` – a degree of dependence of the selected dependent
  - 0 → no dependence
  - 1 → 1. degree of dependence
  - 2 → 2. degree of dependence
  - 3 → 3. degree of dependence
  - 4 → 4. degree of dependence

When these two variables are set, the main phase of generating dependents is finished. After that, DCS generates two tables with the numbers of caregivers and the dependents:

- `caregivers.fac` – the remaining number of possible caregivers according to age and gender, who were not assigned any dependent in the main phase
- `dependants.fac` – the remaining number of possible dependents according to gender, age category and degree of dependence, who were not assigned a caregiver during the main phase

Both these tables are subsequently used in the second phase of generating dependents, which is implemented in DCS `09_SPCODES.DCS`. In this second phase, which is the corrective phase, all

remaining dependents – if possible – get assigned to a main person who weas not assigned care in the first phase

If the number of possible caregivers in table `caregivers.fac`, who are of age and gender corresponding to the main person, is a positive number, the person is assigned as a potential caregiver. Subsequently, it is verified that table `dependants.fac` contains dependents in age and gender which correspond to the partner, child, or parents of a possible caregiver. If at least one such dependent exists, the main person is then definitively labelled as a caregiver and Is assigned this dependent including their degree of dependence. This approach ensures the correct number of caregivers and dependents in MP. Caregivers correspond also to the real distribution by age. In this aspect, there is a larger inaccuracy in case of dependents because they are artificially assigned their year of birth as 1985 for many children and also due to a larger number of dependents from the age group 66+ years assigned closer to the lower boundary of this age group. If in the future, caregivers and corresponding dependents were paired directly, this would mean an interference also in the process of generating partners and children.

6.5.2 Prophet

The implementation of care for a dependent can be divided into 3 parts:

Modelling Care for a Dependent

New statuses were added – `PERSON_CARE` with indication whether there is care or not, and `PERSON_CARE_LEVEL` with the degree of dependence of a dependent.

A dependent can be a partner, parent or a child, minimum age of a dependent is 1 year, one caregiver can only have one dependent at any time. The consequent care is modelled. The minimum age of a caregiver is 15 years of age.

The model can decrease working hours of a caregiver based on the caregiver's age and gender and also based on the degree of dependence of their dependent, currently set at 100 %.

The commencement of care is modelled by event `Person_Care_Start`, probabilities are set based on the main person, the event generates the dependent and the degree of dependence.

The termination of care either happens for a reason of the dependent's passing or by event `Person_Care_Stop` (i.e., termination of care for a reason other than decease).

A change of the degree of dependence is performed by event `Person_Care_Change` in two phases – first, the event of change is generated, then a new degree of dependence is generated.

After adding new stochastic events, it was necessary to increase the value of variable `NO_EVENTS` in table `global.fac` by three.

Care for a dependent is modelled separately from family care, family care was deactivated (i.e., it is not used in the model – `INIT_FAMILY_CARE = 0`).

Modelling Parents Directly in Model Points

Parents are modelled similarly to children in a field with dimension `PARENT` of size 2 – age, gender and status (whether they are alive or have passed) are modelled for each parent.

Modelling Increased Mortality of Dependents

Probabilities of decease of a dependent are increased by the excess-mortality coefficient. This coefficient depends on age and degree of dependency (it can be turned off with a switch in `global.fac` file). The increased mortality rate is then projected in all calculations which depend on a mortality rate.

It was set that dependent children cannot leave the household and thus, mortality is modelled for them for the entire time of their dependence. Similarly, the mortality rate of children is in the model added also to children who are dependent and have left their household in the past.

6.5.3 Description of .fac Tables

mort_care_adjust.fac

This table contains excess-mortality coefficients for dependents.

Table 6.1: Structure of Table mort_care_adjust.fac

Code	Comment
AGE_NOW_Y	Current age of the dependent
LEVEL	Degree of dependence
PERSON_CARE	Excess-mortality coefficient

person_care_change.fac

This table includes probabilities of transitions between individual degrees of dependence.

Table 6.2: Structure of Table person_care_change.fac

Code	Comment
AGE_NOW_Y	Current age of the dependent – category
PER_CARE_LEVEL_OLD	Current degree of dependence
PER_CARE_LEVEL_NEW	New degree of dependence
CHANGE_PROB	Monthly probability of transition

person_care_change_age.fac

This table contains categorisation of age for table `person_care_change.fac` mentioned above. The table must be ordered by age and each entry must be unique.

Table 6.3: Structure of Table person_care_change_age.fac

Code	Comment
CATEGORY	Category (1-3)
MIN_AGE	Lower age boundary of the given category

person_care_init_prob.fac

This table is used for selection of a dependent and the commencement degree of dependency

Table 6.4: Structure of Table person_care_init_prob.fac

Code	Comment
AGE_NOW_Y	Current age of the dependent (from 0 to 100 years)
PER_CARE_LEVEL	Degree of dependence
INIT_PROB	Monthly probability of selection of a dependent and a commencement degree of dependence

person_care_salary.fac

This table includes coefficients of wage reduction during care. It is currently set at 100%.

Table 6.5: Structure of Table person_care_salary.fac

Code	Comment
SEX	Gender of a caregiver
AGE_NOW_Y	Current age of a caregiver (from 15 to 100 years)
PER_CARE_LEVEL	Degree of dependence
PEN_CARE_SAL_RATIO	Coefficients of wage reduction

person_care_start.fac

This table contains probabilities of the commencement of care for a dependent.

Table 6.6: Structure of Table person_care_start.fac

Code	Comment
SEX	Gender of a caregiver
AGE_NOW_Y	Current age of a caregiver (from 15 to 100 years)
PERSON_CARE_START	Yearly probability of commencement of care for a dependent (monthly probabilities are calculated in the model)

person_care_stop.fac

This table contains probabilities of termination of care for a dependent for reasons other than passing of the dependent.

Table 6.7: Structure of Table person_care_stop.fac

Code	Comment
AGE_NOW_Y	Current age of the dependent (from 0 to 100 years)
PER_CARE_LEVEL	Degree of dependence

PERSON_CARE_STOP	Yearly probability of termination of care for a dependent (monthly probabilities are calculated in the model)

prob_parent_age_diff.fac

This table contains probabilities of age differences between the main person and their parents.

Table 6.8: Structure of Table prob_parent_age_diff.fac

Code	Comment
PARENT_AGE_DIFF	Age difference between a child and a parent
COUNT	The number of parents with the given age difference
PROBABILITY	Probability of existence of the given age difference between a child and a parent

life_expect_rates.fac

This table contains probabilities of life expectancy for a given age.

Table 6.9: Structure of Table life_expect_rates.fac

Code	Comment
AGE	A person's age
LIFE_EXPECT_RATE_MALE	Probability of life expectancy of a man
LIFE_EXPECT_RATE_FEMALE	Probability of life expectancy of a woman

prob_person_care.fac

This table contain probabilities that a person of a given age and of a given gender will be caring for a child, a partner or their parents.

Table 6.10: Structure of Table prob_person_care.fac

Code	Comment
AGE	Age of a caregiver
PROB_CHILD_CARE_MALE	Probability of care for a child in case of men
PROB_CHILD_CARE_FEMALE	Probability of care for a child in case of women
PROB_PARENT_CARE_MALE	Probability of care for parents in case of men
PROB_PARENT_CARE_FEMALE	Probability of care for parents in case of women
PROB_PARTNER_CARE_MALE	Probability of care for a partner in case of men
PROB_PARTNER_CARE_MALE	Probability of care for a partner in case of women

person_care_count_age_cat.fac and dependants.fac

This table includes the number of dependents in 3 age categories divided according to the degree of dependence and gender. Data are from the end of year 2019.

Table 6.11: Structure of Table person_care_count_age_cat.fac and dependants.fac

Code	Comment
AGE_CAT	Age category of a dependent
LEVEL_1_M	The number of men in 1. degree of dependence
LEVEL_2_M	The number of men in 2. degree of dependence
LEVEL_3_M	The number of men in 3. degree of dependence
LEVEL_4_M	The number of men in 4. degree of dependence
LEVEL_1_F	The number of women in 1. degree of dependence
LEVEL_2_F	The number of women in 2. degree of dependence
LEVEL_3_F	The number of women in 3. degree of dependence
LEVEL_3_F	The number of women in 4. degree of dependence

caregivers_counts.fac and caregivers.fac

This table contains the number of caregivers divided according to gender and age.

Table 6.12: Structure of Table caregivers_counts.fac and caregivers.fac

Code	Comment
AGE	Age of a person
MALES	The number of male caregivers of a given age
FEMALEL	The number of female caregivers of a given age

6.6. Implementation Feasibility Assessment of Other Factors

The implementation difficulty of an additional factor region is estimated as low / medium. Implementing region into one table is without difficulty, however, it would be likely necessary to be extend most of the above-mentioned tables (perhaps all of them). It is important to distinguish between the addition of factor region for a dependent and for a caregiver. Adding the factor of a dependent person's region would entail extending tables `mort_care_adjust.fac`, `person_care_change.fac`, `person_care_init_prob.fac` and `person_care_stop.fac`. Adding the factor of a caregiver's region would entail extending tables `person_care_salary.fac` and `person_care_start.fac`.

The implementation difficulty of an additional factor of the dependent's gender is estimated as low / medium. Similarly to implementing the regional factor, the implementation of the additional factor of the gender of the dependent into one table is without a difficulty, however, it would be probably necessary to extend all four relevant tables mentioned for a dependent above.

The implementation of additional factors region and gender of the dependent would primarily entail an addition of another column (columns) including information about the additional factors and determining probabilities / coefficients in given tables.

Adding additional factors (if they proved significant in the future or if better data were available) would increase the size of tables multiple times. In case of a potential implementation of additional factors, it is important to be wary of the effect this may have on calculation times.

6.7. Summary and Evaluation of Other Factors

Table 6.13: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Age of a Caregiver	Excellent	High	Implemented	Implemented
Age of a Dependent	Excellent	High	Implemented	Implemented
Gender of a Caregiver	Excellent	High	Implemented	Implemented
Gender of a Dependent	Data is not available	We expect a significant impact	Low / Medium	We recommend implementation after quality data is obtained
Region	Excellent	Low	Low / Medium	We do not recommend implementation due to low significance

7. Chapter 5, Region

7.1. Introduction

Based on the feasibility study (Deloitte, 2014), a decision was made to add factor of region to the micro-simulation model NEMO. The aim of this addition is to reflect geographical variability of the Czech population and to make the model more accurate by doing so.

According to the above-mentioned feasibility study, it is suitable to use information about region in order to improve particularly the following areas of the model:

- Unemployment – high significance – foreign model analysis shows that this factor can have a significant impact especially on probabilities of transitions to (un)employment due to the fact that job offers vary across regions.
- Wages – high significance – it is expected that the salary amount can be highly dependent on the region where a given person lives.
- Death rate – medium significance – based on analysis of foreign models, it is expected that this factor can have a high significance. For example, in industrial areas higher mortality is observed.

Besides the addition of statistical information about a person's region itself, also an event of moving house (which enables individuals to change their residence) was added.

Geographical terms *region* and *locality* are used in the text of this chapter. These terms are defined as follows. The word *locality* is used for differentiation between rural and urban areas – thus, only two localities exist: "urban" (for towns and cities) and "rural" (for villages and countryside). The word *region* stands for a combination of a district and locality, we distinguish altogether 146 regions, for example "Znojmo, city", or "Jindřichův Hradec, rural".

7.2. Data Sources

The main source of data for most analyses was database STATMIN VZ where postal codes (PSČ) were used as information about region. Databases STATMIN VZ OSVČ and STATMIN ANOD were used for sub-analyses and also to assign postal codes to the model point database.

7.3. Determination of Granularity

The basic granularity of a region which was selected was district ("okres" in Czech) with differentiation of locality – either urban or rural, see Figure 7.1. This level of granularity was chosen based on expert judgement and the required level of detail in model NEMO. Contrary to granularity at regional level ("kraj" in Czech), this more detailed classification better reflects differences in salaries in individual regions and it also enables better characterisation of especially large and diverse administrative regions such as for example Central Bohemian Region and South Moravian Region.

Suitability of such granularity was evaluated based on distribution of assessment bases – inner-group variability decreases by division into districts and localities (assessment bases within districts and localities are more similar to one another than within an administrative region) while maintaining variability among individual regions (differences between regions are not lost).

The final number of districts is 77, out of which 4 districts are solely urban and 4 districts are solely rural. The total number of regions with classification of district and urban/rural locality is 146.

7.4. Preparation of Postal Code (PSČ) Codebook

Basic information about region is present in the source databases in the form of postal codes (PSČ), information about region is implemented in the same form also in the model point database (see further). For further analyses and implementation into the model, it is necessary to unambiguously map postal codes to individual districts and assign a locality for each; i.e., to decide whether it is an urban or rural area.

Czech Post (Česká pošta, 2021) database was used as a source database for postal code mapping to districts and localities urban/rural – this database of the Czech Post (Česká pošta) includes postal codes for settlements up to a level of town boroughs. Each postal code from this scope was assigned a district which more rows (towns and boroughs) were falling under in the source dataset.

Majority of postal codes (94,8 %) fall solely under one district. 14 postal codes (0,5 %) had equal number of entries in multiple districts – for those, a district which they were assigned was chosen randomly.

Moreover, information about code NUTS/LAU were added to the postal code entries based on codebook of the Czech Statistical Office (ČSÚ) (ČSÚ, 2021). Also, information about code of District (alternatively Prague) Social Security Administration was added based on codebook of the Czech Social Security Administration (ČSSZ, 2021). Czech Social Security Administration (ČSSZ) code was assigned based on a name of the district, in case of Prague based on a name of a specific borough.

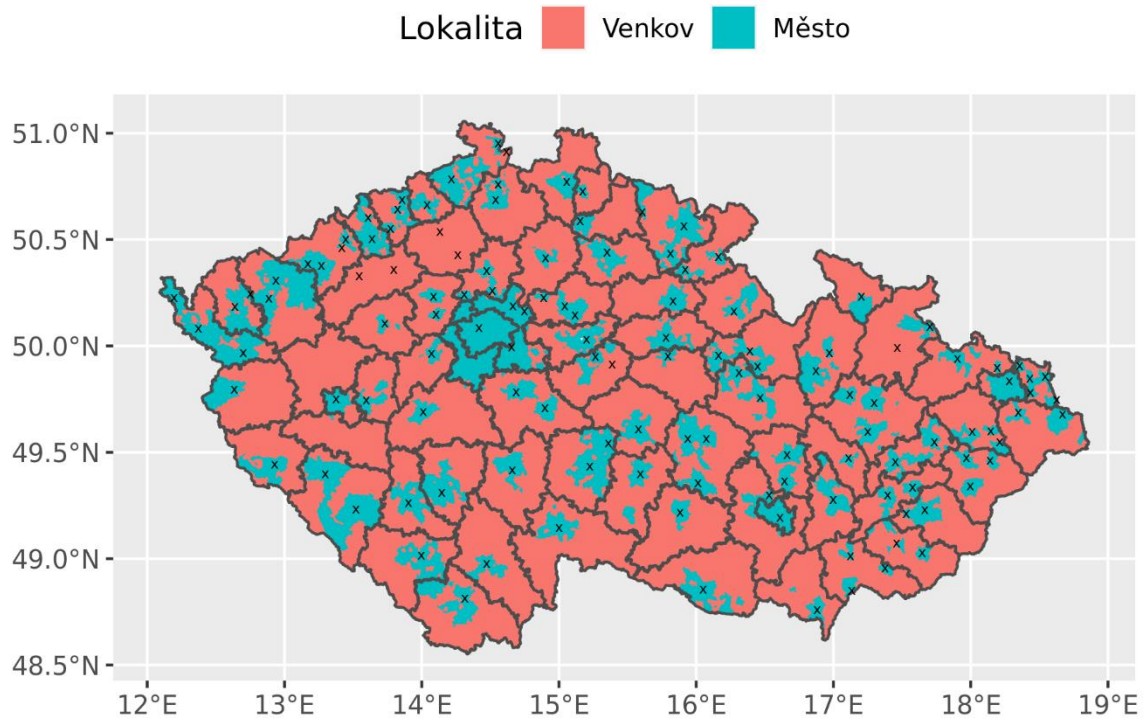

Figure 7.1: Visualisation of postal code classification into localities within districts – urban areas are portrayed red, rural areas are portrayed turquoise. Symbols "x" mark locations of cities with more than 10,000 inhabitants.

Locality (urban/rural) for individual postal codes was determined by the following method. Based on data from the Czech Statistical Office (ČSÚ, 2021), as urban were denoted all towns with population at least 10,000 inhabitants (altogether 129 towns). The postal code for a rural area was determined

as such postal code which at least partially enters one of the towns. Particularly in the Czech Post database, for every row (town or borough) a determination was made whether it belongs to one of the towns (based on value in column *municipality name*). In the next step, a ratio of "urban" rows was calculated for each postal code. As "urban postal code" was marked such postal code for which there were at least 10% rows denoted as urban. For example, postal code 59102 (see Table 7.1) contains altogether 5 rows, 2 of which (40%) are a part of town Žďár nad Sázavou and 3 are not part of any town. Because more than 10% rows are rural (towns), we consider postal code 59102 to be rural.

Table 7.1: An example of determination of urban or rural locality for a given postal code. This postal code is considered to be an urban locality because 40% of rows contain Žďár nad Sázavou under the municipality name and Žďár nad Sázavou is a town (urban area).

Borough Name	Postal Code	Municipality Name
Cikháj	59102	Cikháj
Polnička	59102	Polnička
Světnov	59102	Světnov
Stržanov	59102	Žďár nad Sázavou
Žďár nad Sázavou 2	59102	Žďár nad Sázavou

This methodology selects such postal codes which at least partially enter one of the bigger towns and thus should behave similarly to towns. Above the scope of this methodology, based on expert judgement as urban areas were denoted also the entire district Prague-East and the entire district Prague-West. With these adjustments, we obtain in database STATMIN VZ approximately 4.5 million people living in urban areas and 3.5 million people living in rural area, which approximately corresponds to the real structure of population (52% people live in towns with population above 10,000 inhabitants (ČSÚ, 2021)).

Final classification into districts and localities is displayed in Figure 7.1, which also portrays individual towns which were used as default points for locality determination.

7.5. Impact Analysis of Factors

The aim of the analysis was to determine significance of dependence of moving house event on factors age, gender, current region, marital status and education. Out of these factors, factor age and current region were implemented. The remaining factors were only analysed in regard to their significance and suitability for future implementation into the model.

Factors were analysed from two viewpoints – whether the factor has significant impact on the probability of moving house, and whether it has impact on conditional probability of a new region selection.

All analyses (with the exceptions of factor education) were carried out on data from database STATMIN VZ. The reason for this was especially good availability of detailed data about region during long timeframes (year 2004 – year 2019) and a large amount of contained individuals (approximately 7.4 million unique IDs). On the contrary, database ANOD includes information about region (only district, not locality) only for years 2018 – 2019 - i.e., it includes only a very limited number of individuals who changed their region.

7.5.1 Current region

Data Availability

Current region is present in data in the form of a postal code (STATMIN VZ, STATMIN VZ OSVČ, STATMIN ANOD since year 2019), alternatively in the form of a district (ANOD since year 2018). Since year 2019, there is also information about residence municipality in STATMIN VZ but this piece of information was not used due to too much detail and short history.

Quality of data in all sources is very good, only in database STATMIN VZ the information is absent for 0.2 % of records every year, it is not absent in any other sources.

Information about region was so far missing and has been added (in the form of postal codes) as a part of this project.

Impact Analysis

The impact of a current region on moving house was evaluated as a necessary factor in regard to the event of moving house, especially for determination of the new region. Besides general trends of moving to Prague and other large cities which are popular destination regardless location of current residence, we observed strong tendencies to move to geographically nearby regions.

7.5.2 Age

Data Availability

Information about age is present in all databases in the form of a year of birth, which means that data availability for the analysis is excellent. Since year 2019 database STATMIN VZ also includes a month of birth but this piece of information is not relevant in the given context.

Impact Analysis

An analysis based on data from database STATMIN VZ (Figure 7.2, blue) showed that probability of the event of moving house is strongly dependent on age. The probability of moving house gradually increases with increasing age with a peak around the age of 30. It then decreases with increasing age from that point. Based on the data, a decision was made to divide the model into four age categories: less than 25 years old, 25 – 34 years old, 35 – 50 years old, and over 50 years old.

This analysis also used data from database ANOD for comparing the probabilities of moving house in more mature age. The comparison shows that the probability of moving slightly decreases after retirement, however, the trend remains. The overall trend changes in even higher age, specifically above 80 years of age - this is probably connected to some individuals moving to nursing homes or to their care givers. However, this dependency is not particularly relevant for model NEMO.

Based on Chi-squared test analysis, it was verified that conditional probabilities of selection of a new region differ significantly within each age category in comparison to the population average and differ significantly also among categories.

The factor of age proved to be significant at such a high level, that it was used also during analysis of other factors.

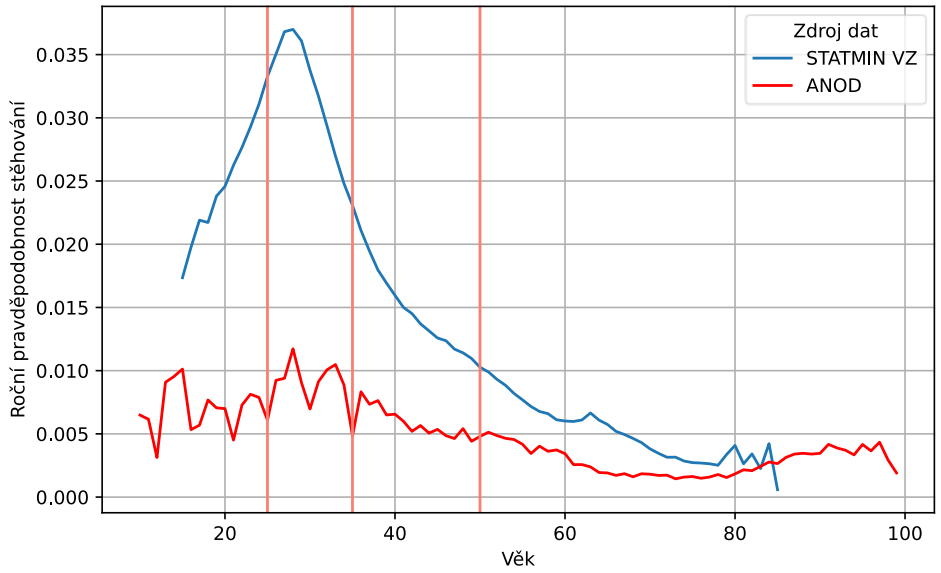

Figure 7.2: Probability of moving house dependent on age – data from database STATMIN VZ (blue) and from database STATMIN ANOD (red), red vertical lines illustrate division of age categories.

7.5.3 Gender

Data Availability

Information about gender is present in good quality in each of the databases which were used.

Impact Analysis

An analysis of impact of gender on the probability of moving house (see Figure 7.3) shows that a difference between men and women occurs especially in a younger age. Women aged 35 and below (first two age categories) move house more often than men, there is a 30% relative difference in probabilities of moving between the genders in the category 0 - 24 years of age and it is 20% in the category 25 - 35 years of age. The difference is not as high in the higher age categories – it appears to rather correspond to an overall shift of the curve for approximately 3 - 5 years towards younger ages for women.

A data analysis from database STATMIN VZ did not prove any dependency on age of probabilities in regard to the new region selection (Chi-squared test at the significance level $\alpha = 5$ %).

We consider gender to be a significant factor for the purpose of determining whether an individual will move house but not for the purpose of determining where they are going to move.

Additionally, due to the fact that it is primary the timing of moving house distribution that is shifted and the trend does not depict any fundamental differences in behaviour, we do not recommend this factor for implementation.

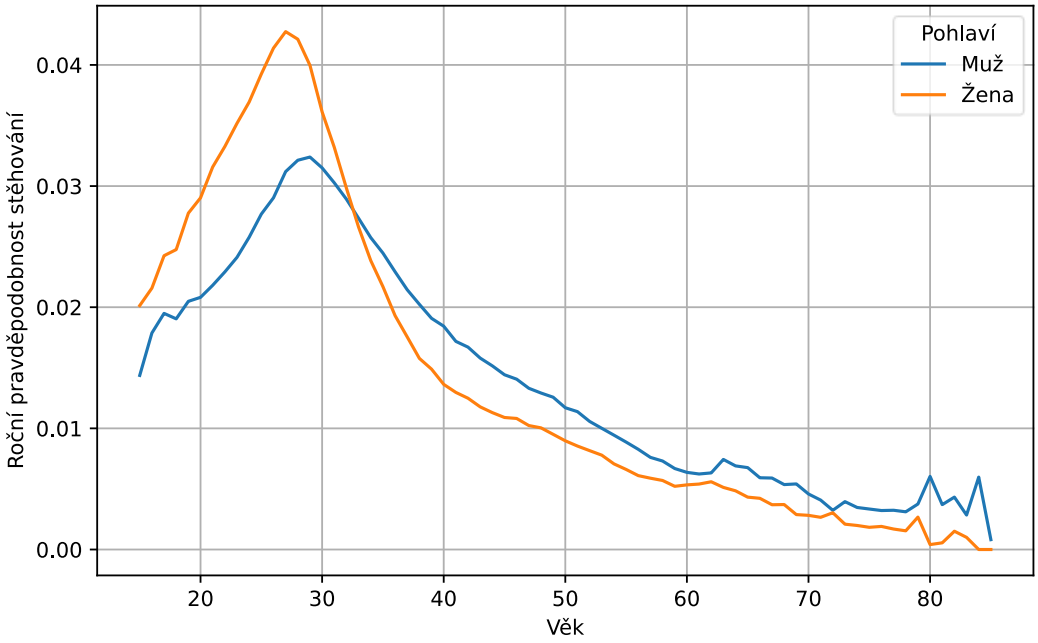

Figure 7.3: Probability of moving house dependent on age and gender.

7.5.4 Education

Data Availability

Information about education (more precisely the highest education achieved) is available in good quality in database „Extended STATMIN VZ“, however, only for a limited number of individuals (approximately 2.5 million unique individuals in comparison to approximately 7.5 million in the basic VZ). Information about education present in the form of a detailed code of the academic institution type, which for our purposes was replaced by the level of education based on classification by the National Institute for Education – „Primary and None“, „Secondary without Completion of Graduation Exams (maturita)” „Secondary with Completion of Graduation Exams (maturita)“ and „University“.

Model points include information about education in lower quality as it was derived based on the age at which the individual entered the labour market.

Impact Analysis

The available data (see Figure 7.4) shows that the only practically significant impact on the probability of moving house is in case of university education and secondary education with completion of graduation exams (maturita) in the age category 25 – 34 years of age. In the remaining categories probabilities of moving are very similar to one another.

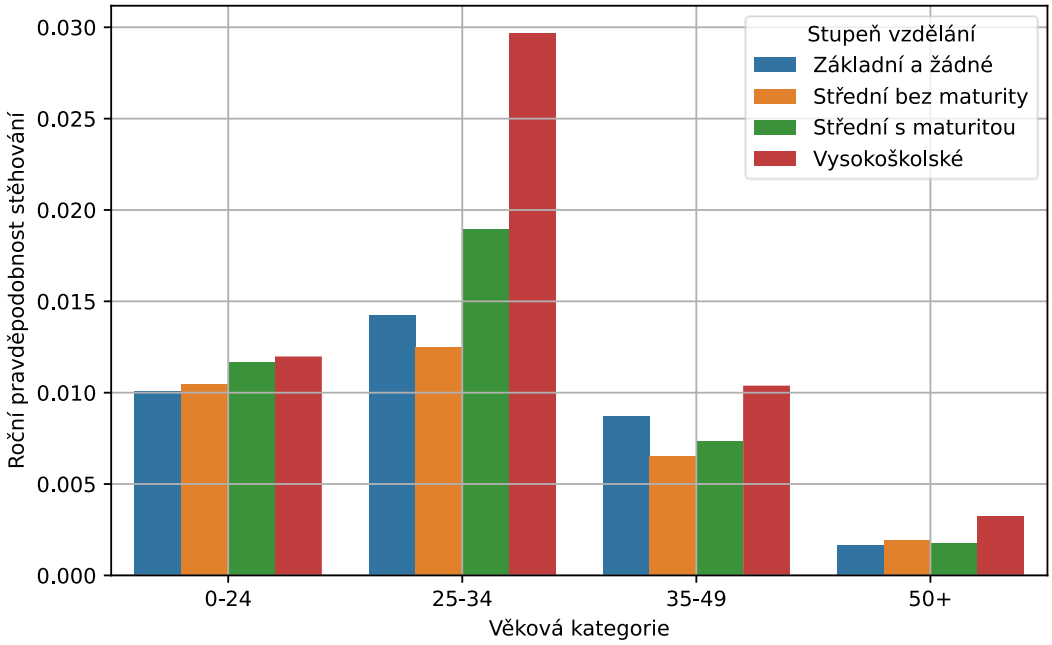

Figure 7.4: Probability of moving house dependent on age category and level of achieved education.

Two districts with high concentration of universities, i.e., with high density of university educated individuals, were analyses in relation to moving house - Prague and Brno-city (Figure 7.5). Both cities show a similar trend of a significant influx of university educated individuals especially around their 30 years of age. No significant trend was apparent in the remaining age categories.

Since mobility of university educated people especially is relatively high, the impact of the factor of education on the overall model can be quite interesting and thus, we recommend this factor at least for consideration at the time of future implementation.

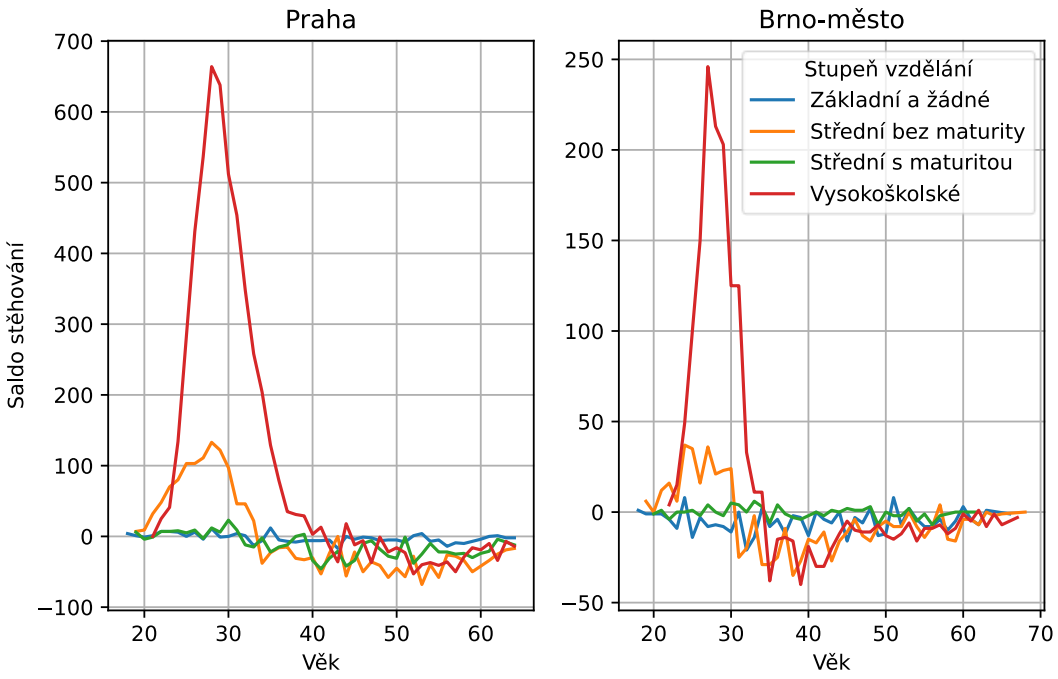

Figure 7.5: Balance of population movement in Prague and Brno-city, according to age and highest achieved level of education.

7.5.5 Marital Status

Data Availability

Information about the number of children for women and about widow(er)s is easily available in database ANOD. Data based on which differentiation of single, divorced or married people could be done is not available, neither is data about the number of children for men.

Such information is in database MP, according to the documentation, however, it was generated stochastically based on occurrence in population, hence it does not carry any actual value for the purposes of this analysis.

Impact Analysis

Due to selection bias caused by availability of data in database ANOD only about retired women and given a relative lack of data about region in this database, this factor was not analysed.

7.6. Implemented Changes

7.6.1 MP Preparation

A column which is newly added to Prophet in regard to this chapter is column `INIT_ZIPCODE`. This new column is first added at the end of database `INEP_PARTICIPANTS`. This happens when starting DCS `MERGE_INEP_EXT_INPUTS`. The input file including people's IDs and a corresponding postal code must be placed into the directory `INPUTS/INEP_EXT_INP_REGION.csv`. Preparation of this input file is described below.

Only a small change in input and output format was carried out in most DCS programmes - in this case it was an addition of input column `INIT_ZIPCODE` into input format `INEP_modelpoints` and its addition into output format `Modelpoints`.

Inner-code changes in DCS `03_newborn_and_children.DCS` and `04_immigrants.DCS` were performed only for newly generated persons (children and immigrants).

When generating new-borns and children, first a probability determining which postal code will be assigned to a given child is selected based on a random number. Table `newborn_zip_code.fac` (see Description of .fac tables) was added in DCS for this step. The selected postal code is then saved as column `INIT_ZIPCODE`.

An analogical change generates postal codes for immigrants. The only difference is that the table used for generating postal for immigrants is table `immigrants_zip_code.fac`.

Assigning region to individuals in model points

It is necessary to determine the current region of each individual in the model point database (table `INEP_PARTICIPANTS`). Ideally, postal code of the given person is used because each postal code corresponds to exactly one of by us invented regions. This choice is also beneficial for future use because it enables an easy change in granularity of the region (each postal code can be assigned one administrative region, district or other administrative unit).

A pairing table (see Preparation of Postal Code (PSČ) Codebook) was prepared in order to assign a district and locality to each postal code.

Information about postal code was added to the model point database based on data from databases `STATMIN VZ`, `STATMIN VZ OSVČ` and `ANOD` which all include the same ID of a person.

The process of finding a postal code itself combines information from all four databases and looks for the last known information about region for each individual. Some databases (`ANOD` and `STATMIN VZ OSVČ`) indicate only a district, not a postal code. The final determination of the postal code is then given by the following rules in a progressive order:

1. Finding last known postal code within each database
2. Finding last known district within relevant databases
3. Comparison of last known postal code and district, only the newest record is kept
4. In case of data equality, the sequence of source databases is as follows:
    a. `ANOD`
    b. `STATMIN VZ`
    c. `STATMIN VZ OSVČ`
5. In case of having information about both postal code and district of an individual:
    a. If information about postal code is newer than information about district, postal code is used
    b. If information about postal code is older but this postal code falls within the given district, postal code is used
    c. If information about postal code is older and it does not fall within a given district, district is used
6. Those individuals for whom the best information is information about their district get assigned a postal code randomly from the given district's postal codes based on distribution of the population in that given district (for example, such postal code in which 20% inhabitants live will be assigned 20% probability.

7. Those individuals for whom we do not have any information about region get assigned a postal code similarly, also by a random selection based on distribution of the population in the whole Czech Republic.

The described method made it possible to assign the most relevant postal code to each individual from database `INEP_PARTICIPANTS`. Particularly, 94.6% individuals were directly assigned a postal code, 2.0% were assigned postal code based on knowledge of their district and the remaining 3.4% were given a random Czech postal code.

The final pairing table ID-PSČ (ID-Postal Code) in CSV format gets connected to model points manually with the use of DCS tool. We recommend the above-mentioned method of assigning postal code to model points to be added into the tool which creates model points (Apache Spark).

7.6.2 Prophet

Performed Changed

A new state variable `REGION` describing the residency district was added into the model. It is apparent that state variable `REGION` is not subject to activation / deactivation, only to change – stochastic event `Region_Change`. After the addition of the new stochastic event, it was necessary to increase the value of variable `NO_EVENTS` in table `Global.fac` by one. Also, two new variables `NO_REGIONS` (which determines the number of districts) a `NO_LOCALITIES` (which determines the number of localities, in this case only urban / rural) were added in table `Global.fac`.

Probabilities of moving house among districts were implemented in table `Region_change.fac`. This probability depends on the current district and age. A new variable `LOCALITY` was also added into the model – this variable determines whether a given person lives in a rural or in an urban area. The stochastic event `Region_Change` (i.e., moving house) first simulates whether a given person will be moving house or not and if they will, it then determines in which district they will move and whether it will be in a rural or urban area. Moving from an urban area to a rural area within one district is not possible. Probabilities of locality change are determined in table `Locality_change.fac`. This probability depends only on the current locality. Probabilities of change among districts and localities are set to be independent in the model.

The current method of locality selection implementation is not optimal, but it was chosen due to the level of difficulty which more accurate solution would entail. The impact of constant probabilities of locality change (regardless the new district) will with time result in equalising ratios of urban and rural inhabitants in all districts; for example, the number of people in urban Prague with rural locality will with time increase. A more suitable approach would be to edit the probabilities of locality selection based on the new district, i.e., new Prague inhabitants would obtain an urban locality with the probability equal to 1; and on the other hand, it would be impossible to obtain a rural locality when moving to rural districts. We recommend consideration of this problem and resolving it for further development of the model.

Table `Zipcodes.fac` which enables a conversion of postal codes to districts and localities was also implemented in the model. It was also necessary to add a new variable `INIT_ZIPCODE` in model points. Prophet allows to set granularity of the region at an arbitrary level – this is done in tables `events.fac`, `global.fac` and `zipcodes.fac`. Locality can also be theoretically set at arbitrary granularity in the same tables.

Modelling of the region (district and locality) is performed only for the main person in the given model point. That means that there is an assumption that the remaining family member live in the same region and move house along with the main person.

Implementation Feasibility Assessment of Other Factors

The difficulty of implementation of additional factors (gender, highest achieved level of education, marital status) should be low. It would primarily entail edits to table `Region_change.fac`, where another column (columns), containing information about gender / education / marital status and also determining the probabilities of moving house in column `CHANGE_PROB`, would have to be added.

We see a potential problem in regard to longer time needed for calculations – the table is already quite large (23 409 entries). Adding all additional factors (if they proved significant in the future or if better-quality data was available) would result in up 32 times larger table size, that is approximately 750 000 entries, which could have a rather significant impact on the calculation time.

7.6.3 Description of .fac Tables

Locality_change.fac

This table determined probabilities of change of locality at the event of moving house (Region_Change).

Table 7.2: Structure of Table Locality_change.fac

Code	Comment
LOCALITY_OLD	Current locality
LOCALITY_NEW	New locality
CHANGE_PROB	Monthly probability of change

Locality_codes.fac

This table is used for conversions of a locality code to a locality name.

Table 7.3: Structure of Table Locality_codes.fac

Code	Comment
LOCALITY	Code of locality
DESCRIPTION	Name of locality (urban / rural)

Region_change.fac

This table determines probabilities of a region change in the event of moving house (Region_Change).

Table 7.4: Structure of Table Region_change.fac

Code	Comment
AGE_NOW_Y	Current age category (according to region_change_age.fac)
REGION_OLD	Current region
REGION_NEW	New region
CHANGE_PROB	Yearly probability of change

Region_codes.fac

This table is used to convert a region code to a region name.

Table 7.5: Structure of Table Region_codes.fac

Code	Comment
REGION	Code of region
DESCRIPTION	Name of region

Region_change_age.fac

This table us used for categorisation of age for the purposes of moving house.

Table 7.6: Structure of Table Region_change_age.fac

Code	Comment
CATEGORY	Code of category
MIN_AGE	Lower boundary of the given age category

Zipcodes.fac

This table is used as a postal code converter which converts postal code (PSČ / ZIP) to a region code and a location code.

Table 7.7: Structure of Table Zipcodes.fac

Code	Comment
ZIPCODE	Postal code (PSČ)
REGION	Code of region
LOCALITY	Code of locality

Newborn_zip_code.fac

This table is used to determine postal codes (PSČ) for new-borns and children based on calculated probability.

Table 7.8: Structure of Table Newborn_zip_code.fac

Code	Comment
INDEX	Row number
PROBABILITY	Probability that a given person belongs to the specific postal code (PSČ)
ZIPCODE	Postal code (PSČ)

Immigrants_zip_code.fac

This table is used to determine postal codes (PSČ) for immigrants based on calculated probability.

Table 7.9: Structure of Table Immigrants_zip_code.fac

Code	Comment
INDEX	Row number
PROBABILITY	Probability that a given person belongs to the specific postal code (PSČ)
ZIPCODE	Postal code (PSČ)

7.7. Summary and Evaluation of Other Factors

Table 7.10: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Current Region	Good in STATMIN VZ; insufficient data in ANOD at the moment	High	Implemented	Implemented
Age	Excellent	High	Implemented	Implemented
Gender	Excellent	Medium	Low	Implementation not recommended
Marital Status	Poor	Low	Low	Implementation not recommended
Education	Good	High	Low	We recommend consideration

8. Chapter 6, Occupation

8.1. Introduction

Occupation is a new factor which further extends the micro-simulation model NEMO. The aim of this factor is to reflect career development of individuals in the simulation and also to use known information about occupation for improvement of other areas of model NEMO. Among areas which, according to a feasibility study, can benefit from information about occupation are especially: income (salary/wage) simulation, decisions on retirement commencement and sickness rate of an individual.

In this chapter, suitable determination of granularity for the occupation factor is described (according to classification CZ-ISCO). Moreover, significance of dependence of occupation change on factors age, gender and education is also determined. It is followed by description of the analysis and calculation of probability of occupation change and by use of specific factors.

8.2. Data Sources

Data for occupation analysis were sourced from database Extended STATMIN VZ (subset of records from database STATMIN VZ with the addition of a few columns from database ISPV including CZ-ISCO code and education). Data from extended VZ comprise approximately 42% unique IDs from the total number of unique IDs which are available in database STATMIN VZ, since year 2013 until up to the present.

While analysing information about occupation (CZ-ISCO code), wrong entries were filtered (other than five-digit codes, codes that include other numerical symbols, codes that do not exist in CZ-ISCO codebook, etc.) Simultaneously, IDs that fall within occupation category 0, i.e., employees of armed forces of whom there are approximately two hundred IDs in the database while the number of unique IDs is in single digits.

The occupation category is commonly recorded outside the Extended STATMIN VZ database and the presence of such information in this database is probably a data error. Since the number of entries in category occupation 0 does not correspond to the factual number of records for occupation in armed forces, this occupation category was deleted from the analysis and also from further calculations.

In order to improve assigning occupations to model points (for which is not known in database Extended STATMIN VZ), the client offered an option to connect a database which contains a part of the RES sentence (ČSÚ, 2018) (especially NACE activity branch and region within the scope of the activity). Such data was not used in the current project due demand on time and technical difficulty of joining it with current data sources. With the use of the RES sentence, it would be possible to assign an activity branch to individual codes of employers and then use this piece of information in order to assign occupations more accurately.

The difficulty of the method comes particularly from impossibility of direct mapping of occupation CZ-ISCO codes to codes of NACE activity branches and also from the fact that it is common that people with multiple occupations work in the same company. Despite these problems, we recommend consideration of the use of this data source in a future project.

8.3. Determination of Granularity for Occupation

Codebook CZ-ISCO (used also at an international level in the form of ISCO occupation code) works with five-digit codes which can be categorised into 5 hierarchical levels of detail for individual occupations,

where 1. level is the most general (first digit in the code) and 5. level is the most detailed (all 5 digits of the code). Codebook CZ-ISCO is available in its current version on webpages of the Czech Statistical Office (ČSÚ, Klasifikace zaměstnání (CZ-ISCO), 2020).

It was necessary to decide which level of occupation granularity will be used in analyses and modelling when using the codebook. In order to determine a suitable level of granularity for occupation, persons who are in individual categories and their occupational branches were first analysed. This analysis helped to identify in which categories there were too few people and where, on the other hand, too many; and also, if the occupation categories are consistent regarding scope of work. Next, homogeneity of monthly assessment bases among individual occupational categories was analysed (e.g., differences in assessment bases for occupations 1, 2, 3, and the like for first-level granularity analysis) and also within each category (e.g., homogeneity among occupations 21, 22, 23, …). Logarithm of assessment base was taken before the outcomes were visualised – this was done in order to reflect differences among salaries within individual occupations.

Figure 8.1 shows an analysis of the number of people and distribution of monthly income for individual occupation categories at the second codebook level.

Occupations from the same category are displayed in the same colour. Comparing distribution of income within each category at level 1 shoes that these broad categories often include occupations with significantly different incomes. The most distinctive cases are in occupation 14 (Management employees in accommodation and restaurant services, 23 (Educational Specialists) or 93 (Auxiliary workers in mining, construction, etc.). Further, more detailed differentiation among different occupations could be achieved with categorisation at third level but only if the number of persons in individual categories was significantly lower and if many categories with negligible representation were obtained. Already level 2 there is only a limited number of people for multiple occupations. Low amount of people then leads to observations of only a negligible number of transitions between occupations, and thus probabilities of these transitions are very unreliable, it is also relatively unreliable to use these poorly represented categories for salary predictions.

Based on a compromise between the above-mentioned influences (number of people versus consistency of income) and based on consultation with the client, granularity which was chosen for occupation factor was level 2 granularity of codebook CZISCO which enables to observe occupations and take into consideration individual professional fields, and it is also characterised by a sufficient homogeneity of salaries and a sufficient number of persons in most occupations. This level includes altogether 40 unique occupations – their code comprises the first two digits of the five-digit CZISCO code. Subsequently, for further use of occupations in the model, numbering was created which numbers these categories in a continuous sequence from 1 to 40 with the label `POVOLANI_ID.`

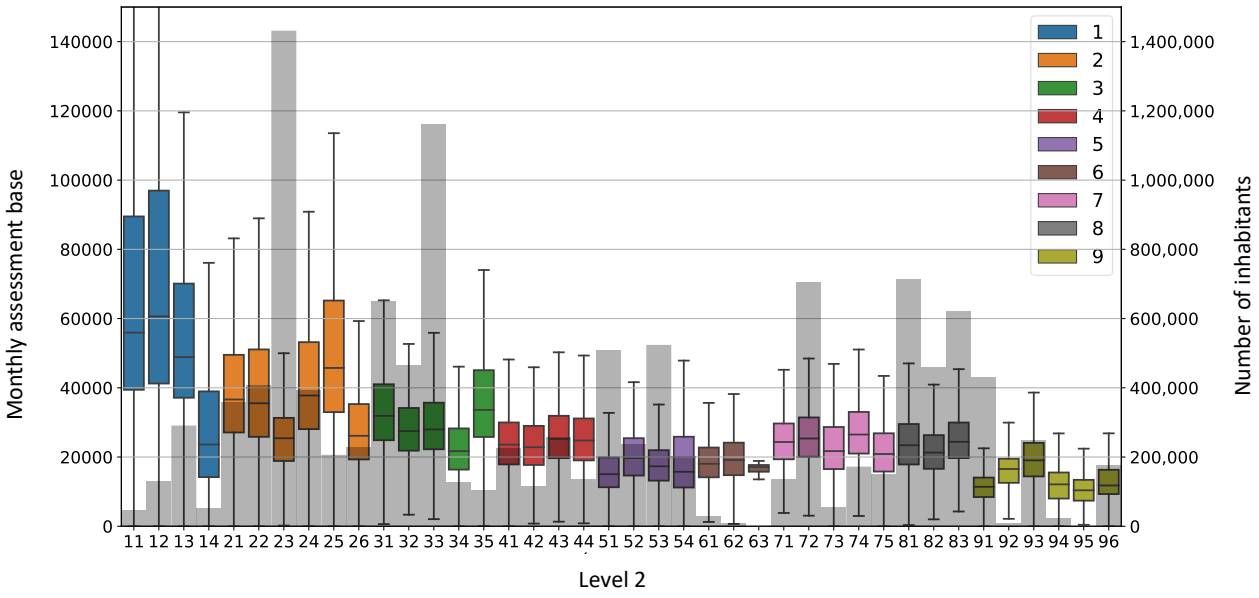

Figure 8.1: Boxplot diagram displaying the median monthly assessment bases logarithms for occupations in granularity at level 2 (left axis y) and the number of persons in each occupation (right axis y).

8.4. Factors Impacting Occupation

8.4.1 Age

Data Availability

Age in the form of year of birth is a factor available for all IDs (persons) in all data sources.

Factor Impact

Probability of occupation change decreases with increasing age (see Figure 8.2). An exception is the time period at the beginning of people's careers approximately until the age of 25 years and also a slight deviation around the time of retirement commencement. Monthly probability of occupation changes without leaving the labour market (left panel) and conditional probability of occupation change upon return from inactivity both decrease with increasing age.

A deviation around 50 years of age for both genders is probably an artefact of data analysis. A deviation for men in 63 years of age is caused by slightly higher probability of occupation change in the time period around retirement commencement.

Age factor was implemented as a categorical variable in categories 0-34, 35-49 a 50+ years, divisions between categories are illustrated in Figure 8.2 by dashed lines.

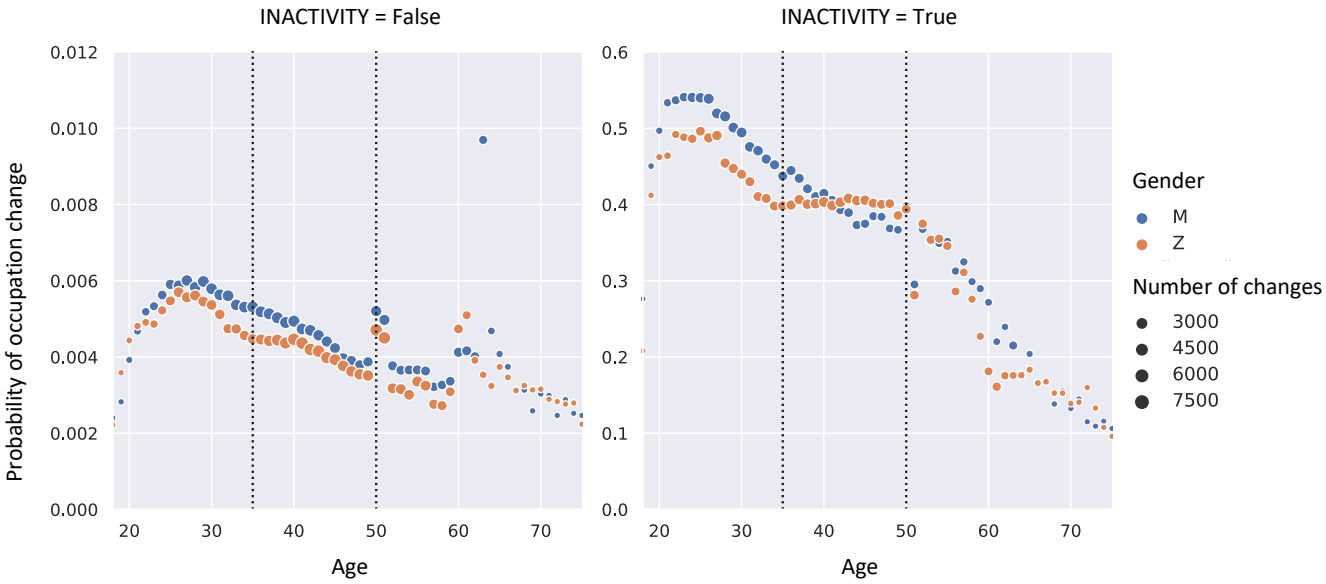

Figure 8.2: Probability of occupation change dependent on age and gender. The size of displayed points corresponds to the number of observed occupation changes. The left panel describes situation when a person changes their occupation without leaving the labour market (monthly probability) and on the right (one-off conditional probability) it is a probability of occupation change upon return to the labour market from inactivity. Vertical axes of both panels use substantially different scales. Dotted lines mark boundaries of intervals which were used for implementation of tables in Prophet.

8.4.2 Inactivity

Data Availability

The aim of this factor is to reflect various behaviour of people who return to the labour market after a period of inactivity and when doing so they change their occupation. Factor of inactivity was added in the model after consultation with the client.

We define inactivity as an interruption of gainful activity for at least one month. Upon return from inactivity (i.e., first month which follows interruption of gainful activity) we evaluate whether the individual changed their occupation compared to their original profession.

Availability of data is excellent, the factor is derived from information about the start and the end of employment.

Factor Impact

Occupation change after a period of inactivity is a relatively common phenomenon – the amount of people who change their occupation upon return to the labour market is in tens of percent (see Illustration Figure 8.2), this value decreases with age. This probability is approximately 50% for younger people and it decreases with age up to approximately 20% at pre-retirement age.

Comparison with probability of occupation change without leaving the labour market is difficult because in one case (while remaining on labour market) it is a monthly unconditional probability which gets applied repeatedly during the individual's lifetime, while in case of occupation change after inactivity it is a probability conditioned by the time period of inactivity and thus applied once. Approximately 60% individuals who change their occupation do so without leaving the labour market and the remaining 40% of occupation changes happens after returning from inactivity. This ratio is

dependent on gender only slightly, with increasing age there are more occupational changes without leaving the labour market, fewer after commencement of retirement.

Factor of inactivity was evaluated as significant and was implemented.

8.4.3 Gender

Data Availability

Gender is another basic factor which was used during analysis and probability calculations. Information about gender is available for each personal ID in all data sources.

Factor Impact

Probability of occupation change dependent on gender factor plays a certain role, especially in old-age pension age (see Figure 8.2). Monthly probability of occupation change is slightly higher for women (especially while in age of gainful activity). In contrast, monthly probability of occupation change for men increases with increasing age. This factor is implemented in probability calculations.

8.4.4 Education

Data Availability

Information about education is available for all rows of database Extended STATMIN VZ in the form of a detailed code of educational institution. For the purposes of joining with model NEMO, this code was replaced by an achieved level of education based on classification of the National Institute for Education – "Primary and None" (Národní ústav pro vzdělávání, 2021), "Secondary without Completion of Graduation Exams (Maturita)", "Secondary with Completion of Graduation Exams (Maturita)" and "University".

Factor Impact

The factor of education has a significant impact especially on the choice of occupation generally speaking (see Figure 8.3). In contrast, the impact of it on occupation change is relatively small.

Based on the structure of educational system, we can divide occupation into three or four large groups that show similar trends.

The first group is occupation 11-26 where university education prevails. This group includes category 1 (Lawmakers and Management) and category 2 (Specialists). The only occupation which deviates from the trend is occupation 14 (Management employees in accommodation and restaurant services, sales), where the ratio of university educated people is significantly lower.

The second large group is occupation 31-44 where prevailing education is secondary education with completion of graduation exams (maturita). It comprises of occupation branch "Technical and Specialist Workers" and "Clerks". A rather unusual occupation here is occupation 32 (Specialist workers in the health care) where lower levels of education are practically absent.

The third large group comprises of all other occupations (51-96) where the prevailing education is secondary education without completion of graduation exams (maturita). Among other types, secondary specialist trade education with obtaining an apprenticeship certificate falls within this category, which leads to mostly to occupations focused on manual work. There is only a negligible number of people with university degree in this occupational group. Slightly different is group 51-54 where is higher representation of secondary education with completion of graduation exams

(maturita), these are Employees in services and in sales. Another two occupations (63 and 74) deviate by comprising of a negligible number of people with primary education. For occupation 63 (Farmers, hunters, …), this is an artefact caused by lack of data, in case of occupation 74 (Electrotechnicians) this is probably given the fact that to perform electrical work it is necessary to have relevant qualification (so-called ordinance 50) (ČR, 2021).

A subset in manual work occupations is group occupation 91-96 which is characteristic by a higher ratio of people with primary education. The occupation type here is "Auxiliary and unqualified workers".

The dependence of probability of occupation change on education is very low and inconsistent. One of the identified trends is the fact that mostly in younger age the most frequent occupation changes are by people with lower (i.e., primary) education.

The impact of education is primarily on the selection of the actual occupation, not on the probability of occupation change. This fact is confirmed also by the structure of transition matrices where frequent transitions are observed within the individual above-described structures. Even without including the factor of education in the transition matrices, it should not happen often that a person with low level of education will work at a highly qualified position. For this reason, we rather don't recommend implementing the factor of education in the process of occupation change.

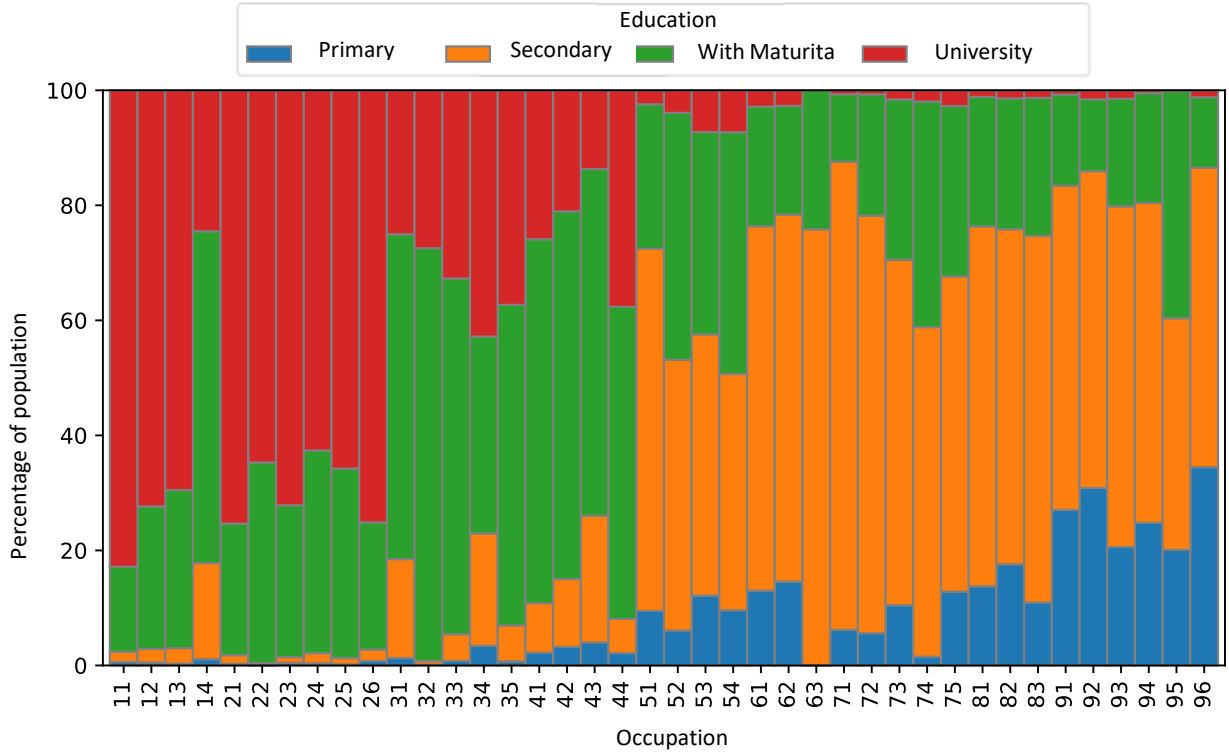

Figure 8.3: Structure of occupations according to current education of gainfully active individuals.

8.5. Calculation of Occupation Change Probabilities

The expansion of database Extended STATMIN VZ into monthly granularity was used for calculation of probability of occupation change in dependence on the above analysed factors (gender, age category, inactivity). The final table includes a unique combination of ID-month records in every line.

It is evaluated for each month in history of a given individual whether they changed their occupation in the following month. The number of these changes (for a group of persons with a given combination of factors) comprise a numerator of the final probability. The denominator is the number of observed ID-month records, i.e., rows in the prepared table. This procedure gives us basic monthly probability of occupation change.

The next step is changing basic probabilities by smoothing. For a negligible part of combination of factors and occupation it was observed too few transitions in the data, in some cases none at all. Probabilities calculated from such limited data do not carry much actual value and are unstable. That is why we perform two-step smoothing – first of "diagonal" probability that a person will not change their occupation, ten "off-diagonal" probability of transitions into a different occupation. All calculations are carried out separately for each combination of factors.

In the first step of changes in probabilities, we alter the "diagonal" probabilities of transitions (p_{ii}), i.e., probability of staying in the same occupation also in the following month – we do so by the means of Bayesian average (see the formula):

$$p_{ii} = \frac{C \cdot p_0 \ + \ x_{ii}}{C + \sum_{j=1}^{N} x_{ij}},$$

where C, p_0 are constants (see further), x_{ii} is the observed number of "transitions" into the same occupation and $\sum_{j=1}^{N} x_{ij}$ is the number of all ID-month records spent in the given occupation. The Bayesian alteration then corresponds to enrichment of the real observation with *C* of other data points for which value p_0 is equal to the average value across the dataset. This method practically does not change the value of probability for points with a high number of observations, and it brings the value closer to the overall average for points with a low number of observations. Different values of constants were used for each combination of factors – as value of *C* the value of a tenth percentile of the number of transitions was used, for p_0 average probability for a given combination of factors was used.

In the next step, the remaining ("off-diagonal") probabilities of transitions are altered in such a way that gives the sum of probabilities for every default occupation equal to one. We use the method of additive smoothing – we first add value 0.5 to every number of transition observations – that gives us non-zero probabilities also for transitions which we were not observed in the data and at the same time we are not introducing aby significant distortion because we are adding the same small value to each element.

$$\mathrm{p_{ij}} = \frac{0{,}5 + \mathrm{x_{ij}}}{\mathrm{N} \cdot 0{,}5 + \sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{x_{ik}} - \ \mathrm{x_{ii}}} \cdot (1 - \mathrm{p_{ii}})$$

The final probabilities are non-zero for each combination of factors, default and new occupation and the number of probabilities for every combination of factors in each default occupation is equal to one.

8.6. Implemented Changes

8.6.1 Preparation of MPs

It is necessary to determine a default occupation for each individual in the model point database. We add the entire five-digit CZISCO code directly in the model points, this code is then converted by a codebook to a two-digit code of occupation category. This method enables potential changes in granularity of occupation or a change in definition of categories. Assigning a category to CZISCO codes is realised with the use of a conversion table. Occupation is assigned to all individuals including retired persons, gainfully inactive person and self-employed people – this is done in such a manner that allows potential transition of the individuals to employment.

The real occupation is known from database Extended STATMIN VZ only for 32% model points. Selection of occupation for the remaining individuals must be done stochastically. The final distribution of occupations for model points should correspond to the real distribution in the population, according to the client's requirements taken from database ISPV (TREXIMA, 2019). Also, the distribution of persons in individual categories of occupations should be consistent with data from extended VZ – for example category 23 ("Educational Specialists") includes especially university-educated women with above-average salaries. In contrast, for example category 72 ("Metalworkers and machinery workers") includes men with lower education (without completion of maturita graduation exams) and with approximately average salaries.

Determination of occupation consists of two basic steps – first, category occupation is stochastically determined based on gender, education and salary, then CZISCO code is determined within each category in such a way that gives similar distribution of codes as in data from extended VZ.

To determine occupation category, we first calculate the number of individuals in each combination of factors (gender, education, salary decile) in extended VZ, we use the last known occupation and education for each person. We calculate probability of a given occupation category for each combination of factor (i.e., we calculate "What is the probability that a randomly selected woman with secondary education with completed maturita graduation exams who receives salary at the level of sixth decile will perform occupation 34?). Here we again use additive smoothing (see above) with constant 0.5 to ensure that no combination of factors and occupation will have a zero probability. We then multiply these probabilities by the number of model points in each combination of factors. This gives us the number of model points in each category with the correct distribution of gender, education and salary both in total (according to model points) and within individual occupations (according to extended VZ).

We calculate the desirable final number of model points in each occupation based on occupational distribution from ISPV and the total number of individuals in model points. We then compare this number with the total number of model points in each occupation which we obtained in the previous step (including differentiation according to factors). We then multiply the number of people in each combination of factors within each occupational category by a correction coefficient in such a way that the final number of model points in occupation corresponds to ISPV. By doing so we obtain the number of model points in each combination of factors and occupation in a way that the total number of individuals in each occupation corresponds to data from ISPV and at the same time distribution of gender, education and salaries within each occupation corresponds to data from extended VZ.

In the next step we take the number of individuals for whom we know the given occupation with certainty from extended VZ and we deduct it from the final number of model points in each

combination of factors and occupation. This gives us the number of model points in each combination of factors for which an occupation must be randomly assigned. In multiple cases it happens that the number of people from extended VZ with a given combination of factors and occupation exceed the desirable number of model points – in such a case, this occupation does not get assigned to any more model points with the given combination of factors. We derive the probability of assigning an occupation for each combination of factors from the number of missing model points and based on these probabilities we stochastically select an occupational category for everyone for whom we did not know the occupation based on extended VZ.

The last step is assigning a five-digit CZISCO code based on a two-digit category determined earlier. For that, probabilities derived from the number of observed individuals with the given code within individual categories is used (based on data from extended VZ), now without differentiation of factors.

The final outcome is a csv file with columns ID and CZISCO code.

CZISCO code enters Prophet as column `INIT_CZISCO`. This new column is first added at the end of database `INEP_PARTICIPANTS`. This happens upon starting DCS `MERGE_INEP_EXT_INPUTS`. Input file including personal ID and a corresponding CZISCO code must be placed in directory `INPUTS/INEP_EXT_INP_POVOLANI.csv`. Preparation of this input file is described above.

In case of most DCS programmes, the changes that occurred were only in the input and output format. In this case specifically adding column `INIT_CZISCO` into input format `INEP_modelpoints` and its addition into the outcome format `Modelpoints`.

The only inner-code changes which occurred were for newly generated persons (children and immigrants) – these changes were carried out in DCS `03_newborn_and_children.DCS` and `04_immigrants.DCS`.

When generating new-borns and children, first probability, according to which CZISCO code is determined for the given child, is selected based on a random number - table `newborn_czisco.fac` (see Description of .fac tables) was added in this step. The selected CZISCO code is then saved as column `INIT_CZISCO`.

The same change generates CZISCO code for immigrants. The only difference is that a table used for generating this code for immigrants is table `immigrants_czisco.fac`.

8.6.2 Prophet

A new state variable `OCCUPATION` determining a person's occupation was added in the model. It is apparent that state variable `OCCUPATION` is not subject to activation / deactivation, only to change – stochastic event `Occupation_Change`. After the addition of the new stochastic event, it was necessary to increase the value of variable `NO_EVENTS` in table `Global.fac` by one. Also, a new variable `NO_OCCUPATIONS` (which determines the number of occupations) was added in table `Global.fac`. State variable `OCCUPATION` gets numbered according to selected granularity (currently 1 - 40). Occupation is the same for employees and self-employed.

A change in occupation can occur during employment (without an event) or during a transition from inactivity (with an event), which happens with different probabilities.

Stochastic event `Occupation_Change` (i.e., change of occupation) first simulates whether a given person will be changing their occupation and if so, a new occupation is determined. Probabilities of

transitions between occupations are implemented in table `Occupation_change.fac`. This probability depends on the current occupation, categorised age and gender.

Description of the individual occupation categories can be found in table `Occupation_codes.fac`.

8.6.3 Description of .fac Tables

CZISCO.fac

This table is used for conversion of CZISCO code to an occupation category.

Table 1: Structure of Table CZISCO.fac

Code	Comment
CZISCO	CZISCO code
OCCUPATION	Category occupation

Occupation_change.fac

This table determines probability of transition between occupations in dependence on age and gender.

Table 2: Structure of Table Occupation_change.fac

Code	Comment
SEX	Gender
AGE_NOW_Y	Categorised person's age
OCCUPATION_OLD	Current occupation
OCCUPATION_NEW	New occupation
CHANGE_PROB	Monthly probability of occupation change
CHANGE_PROB_NO_EVENT	Monthly probability of occupation change with no event

Occupation_change_age.fac

This table is used for categorisation of age for table table `Occupation_change.fac` (described above). The table must in ordered by age and each entry must be unique.

Table 3: Structure of Table Occupation_change_age.fac

Code	Comment
CATEGORY	Age category
MIN_AGE	Lower boundary of the given age category

Occupation_codes.fac

This table contains the code of occupation according to second-level CZISCO codebook and description of individual occupation categories.

Table 4: Structure of Table Occupation_codes.fac

Code	Comment
OCCUPATION	Category of occupation
CZISCO_2	Two-digit CZISCO code
DESCRIPTION	Description of the given occupation category

newborn_czisco.fac a immigrants_czisco.fac

Both tables have the same structure and include probabilities for each five-digit occupation CZISCO code assigned to new-borns and immigrants.

Table 5: Structure of Tables newborn_czisco.fac and immigrants_czisco.fac

Code	Comment
INDEX	Auxiliary column which includes the numbers of rows
CZISCO	Five-digit CZISCO code
PROBABILITY	Probability for a given CZISCO code

8.7. Implementation Feasibility Assessment of Other Factors

Implementation difficulty of additional factor of education should be low. It would entail primarily edits in table `Occupation_change.fac` where another column with information about education would have to be added.

We see a potential problem regarding longer time needed for calculations – the table is already quite large. Adding the additional factor of education could result in multiple times larger table size. In case of future implementation, it is important to be wary of the impact on the calculation time.

8.8. Suitability of Future Implementation of Additional Factors

Table 6: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Age	Excellent	High – with increasing age the probability of occupation change decreases	Implemented with division into 3 categories	Implemented
Inactivity	Excellent	High – upon return onto the labour market people change their occupation in tens of percent of cases	Implemented	Implemented
Gender	Excellent	Medium – probability of occupation change slightly differs	Implemented	Implemented
Education	Insufficient, available data are for 32% model points and 40% individuals from VZ STATMIN	Medium – education impacts selection of occupation but not the probability of occupation change	Low	We rather do not recommend implementation of this factor

9. Chapter 7, Wage

9.1. Introduction

The aim of this chapter is the implementation of a new method of simulation of wages during individuals' lifetimes with the use of a wage equation derived from American model CBOLT (Congressional Budget Office, 2013).

Compared to the original solution, the new solution always uses stochastic modelling of wages, a calculation of income for self-employed persons is implemented independently from the salaries of employees. Newly, a calculation with the use of the salary equation is also used for students. The code of the model was altogether simplified.

The income which is modelled for employees in most cases corresponds to real gross salary (income calculated with the salary equation is increased to the level of minimum wage where needed, this value is then further altered for students and disabled persons). For self-employed persons, assessment base is modelled (earned income decreased by expenditures, with a subsequent alteration for students and disabled persons).

9.2. Data Sources

Database Extended STATMIN VZ is the main source of data for a data analysis and for a creation of the salary equation for employees in this chapter. The data were used because they include information about education and about the occupation code, both of which are important factors in the salary equation model. The database includes records from year 2011 until year 2020. Other data sources are various codebooks, see further in the text.

The data source for preparation of data for self-employed persons is database STATMIN VZ OSVC – this database, contrary to STATMIN VZ database, does not include information about education and the occupation code, therefore there are significantly fewer factors included in the model for self-employed persons.

We will now describe the preparation of data for employees, description of preparation of data for self-employed persons follows.

9.2.1 Preparation and Database Cleansing: Employees

The following information was selected from the Extended STATMIN VZ database:

- Years 2012 to 2019 (incl.), because the database is not complete for years 2011 and 2020.
- Entries for a non-zero assessment base.
- People aged 15 to 75 years (incl.), because younger persons cannot be employed, and older people are employed rarely.
- Only employee relationships based on employment agreements, i.e., first digits of the gainful activity code (KVČ) equal to 1-9, because we want to model salaries only for employees.
- Only persons for whom both postal code information and information about their occupation code exist because both Region and Occupation are important factors for salary modelling. (If any person's record had a missing entry for occupation or postal code, the relevant information was first added from a different record of the same person in the time closest to the record in question.)
- Only rows with a valid postal code (PSČ).

- All occupations except for soldiers, i.e., all records with a non-zero first digit of code CZISCO.

The table includes information about an assessment base and time from (`OD`) and to (`DO`) the given person was working, it also includes information about an excluded period (`VDOBA`). The total number of days in employment is calculated with the formula DO – OD – VDOBA. A daily assessment base (`VZ_DENNI`) is calculated with the formula VZ / (DO – OD – VDOBA).

The data is divided into months with the use of columns `OD`, `DO`. The data is further aggregated in such a way that for each person and each month there is a unique record including information about the total daily assessment base, gender, year of birth, postal code, occupation and education. Assessment bases were converted from nominal to real values to the year 2019 – this was done with the use of salary inflation coefficients (Coefficients of increase in the general assessment base valid for pensions awarded in year 2019).

Then, the total real monthly assessment base is calculated as a multiplication of the daily VZ and an average number of days in one month, which corresponds to number 365.25/12. To remove effects of extreme values, 0.3% of records with the highest assessment base and a 0.3% of records with the lowest assessment base were excluded from the calculation. Age is then added using the following entries: a year the record was created, a birth year of the individual and a column of the region which corresponds to the interaction of a district with a locality (distinction between an urban and rural area).

9.2.2 Preparation and Database Cleansing: Self-Employed

The first step of data preparation is database cleansing.

Only the following records were selected from database STATMIN VZ OSVC:

- People aged 15 to 75 years (incl.), because younger people cannot become self-employed and older people only rarely do.
- Entries for a non-zero assessment base.
- The number of insured days (column `DNY`) equals to the difference between the beginning and the end of insurance (columns `OD` and `DO`).
- People who in one year have exactly one record with a non-zero assessment base.
- Only rows with a valid postal code (PSČ).

Information about a district and a locality was added from an in-advance prepared codebook (see chapter Region).

Next, a tax base is calculated from the assessment base. A daily assessment base is calculated as a ration of the total assessment base and the duration of insurance expressed in days which is decreased by a replacement insurance time. This value is then converted to a monthly value using multiplication by coefficient 365,25/12 (the average number of days in one month) and it is subsequently amended by a wage inflation coefficient. The value is then multiplied by two, which finally gives the tax base.

The distribution of a tax base that was calculated as described above is significantly distorted by the presence of a minimum assessment base (and to a smaller extend also by a maximum assessment base), in addition, its value changes yearly as it depends on the average income of the employee; after an adjustment to reflect a wage inflation this was between 8000 and 8700 CZK in the recent years. The aim of the prediction is to obtain a tax base without the dependence on the current minimum assessment base, therefore the distribution of tax bases had to be adjusted in relation to this dependence.

The correction of tax base distribution (see Figure 9.1, which is referred to further in this section) is based on several assumptions. We assume that the distribution is log-normal distribution, similarly to employees' incomes. We assume that the assessment bases that are higher than the minimum assessment base (at least 2500 CZK higher, blue field on the right side of the illustration) and that are at the same time lower than the maximum assessment base are not influenced by the existence of the assessment base minimum and that they reflect the real tax base distribution. We assume that assessment bases that are under that minimum assessment base value (at least 1000 CZK lower, blue field on the left side of the illustration) also correspond to the real tax bases because these values are of self-employed persons who engage in self-employment as a secondary gainful activity and their minimum assessment base is significantly lower (approximately 3500 CZK lower).

Parameters of log-normal distribution which the data comes from are based on the distribution of assessment bases higher than the minimum. These parameters (an average and a standard deviation) are estimated separately for each gender and age group (in decades) in such manner that maintains a potential dependence of incomes on gender and age. To avoid effects of extreme values, 0.3% of records with the highest tax base were excluded. This distribution is then further corrected (after a consultation with the client) to get a monthly tax base higher than 5000 CZK for all self-employed persons (red dotted line). The final distribution is used to determine a tax base for those self-employed persons whose assessment base was nearing the minimum (and the maximum) assessment base (grey field in the illustration). The final distribution of tax base for self-employed persons (yellow curve) approximately corresponds to data on tax base from the Financial Administration's system ADIS.

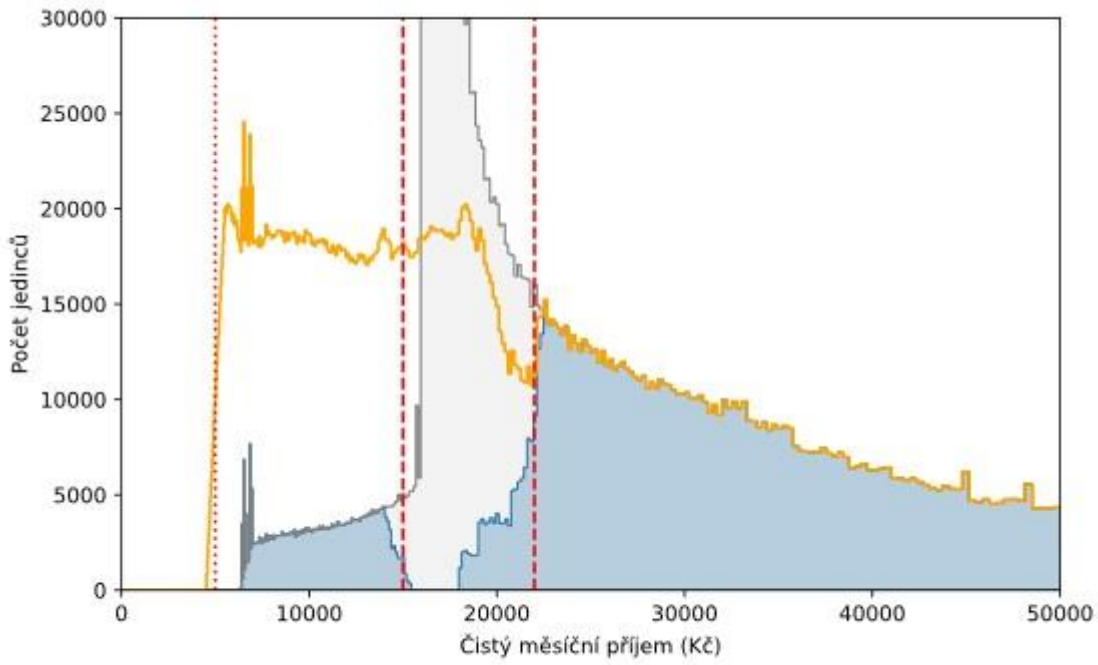

Figure 9.1: Histogram of net monthly income distribution of self-employed persons. Unchanged records are marked blue, changed records are grey. The final distribution corresponds to the area under the yellow curve. Red dashed lines mark boundaries which were used for the exclusion of values effected by a minimum assessment base. The red dotted line represents an artificial minimum of net income.

9.3. Factor Impact Analysis

We analysed the following factors. Selected factors were implemented into a regressive equation which is a part of the wage equation (see the following section). Primarily database Extended STATMIN VZ was used.

9.3.1 Gender

Data Availability

Information about gender is available in good quality in database Extended STATMIN VZ.

Factor Impact

Based on analyses of data for men and women, this factor shows to be very significant. Average values of assessment bases differ, and so does for example the average value of assessment base in relation to age (see Figure 9.2) and to other factors. That means that a so-called interaction between genders and other factors is apparent. For those reasons, we estimate coefficients of linear regression separately for men and women, and thus the factor of gender is not included in the regression models.

9.3.2 Age

Data Availability

Information about age is available in excellent quality directly in the Extended STATMIN VZ as information about a year of birth and a year when the record was created are available.

Factor Impact

This factor is very significant predictor of the assessment base (see Figure 9.2). Since the value of an assessment base increases at the beginning of working life and declines at retirement age, besides age itself, also a second power of age was used – it contributes to capturing the relationship between the assessment base and age more accurately. The factor of age is significant and it is available in good quality, which is why it is included in the regression.

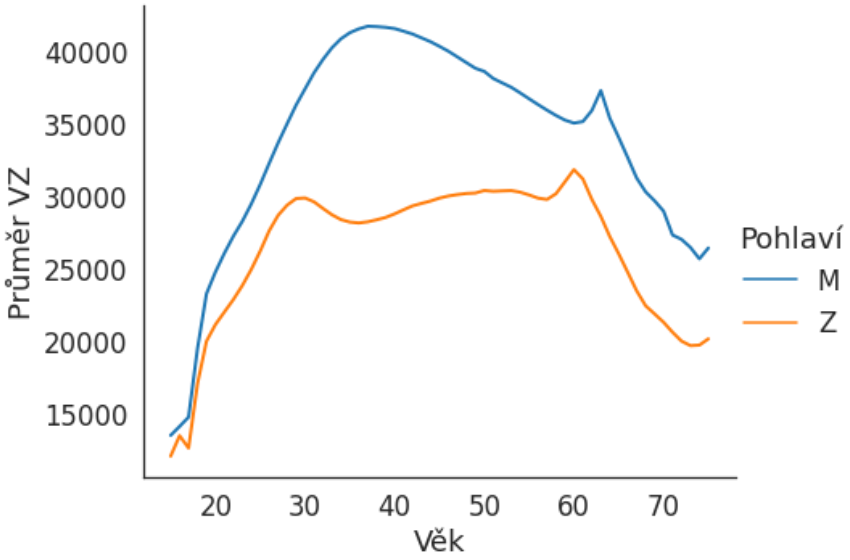

Figure 9.2: Average value of assessment base depending on age and gender. The curve is increasing for men until approximately 40 years of age. In comparison, the average assessment base for women

increases until below 30 years of age, then it decreases (mostly due to taking maternity leave) and then it slightly increases again.

9.3.3 Total Duration of Employment

Data Availability

Database INEP was used to analyse the total duration of employment. The total duration of employment was calculated for every person and year using the number of included years. This information is available in database INEP in excellent quality. However, to be able to include this factor in the model, it was necessary to use database Extended STATMIN VZ which contains key factors occupation and education. Thus, we connected these two databases and added information where the factor could be calculated.

Factor Impact

The factor is correlated with age, and thus it is not possible to use both factors simultaneously. Foreign models use a maximum one of these two factors, see the Feasibility Study. Some foreign models give preference to age (England, USA, Netherlands), on the other hand, others use the total duration of employment instead (France, Sweden).

After adding the factor to the linear regression model and comparing metrics R^2 and RMSE (root mean square error, i.e., the root of an average square error), adding only age versus only the total duration of employment give similar results. When only age is added, the mentioned metrics have better results for men and when only total duration of employment is added, the mentioned metrics have better results for women (probably due to maternity leave), the difference, however, only shows at the third decimal place.

Due to a slightly lower availability of data for total duration of employment compared to data for the factor of age, we recommend using only the factor of age.

9.3.4 Duration of Current Employment

Data Availability

The duration of current employment can be calculated by the sum of consecutive years in database STATMIN VZ. Database INEP cannot be used because it does not include information about employer. The calculated value, however, does not represent the duration of the current employment for persons who have worked in the same employment for more than 16 years (see Figure 9.3), we do not recommend this factor for implementation due to this data limitation.

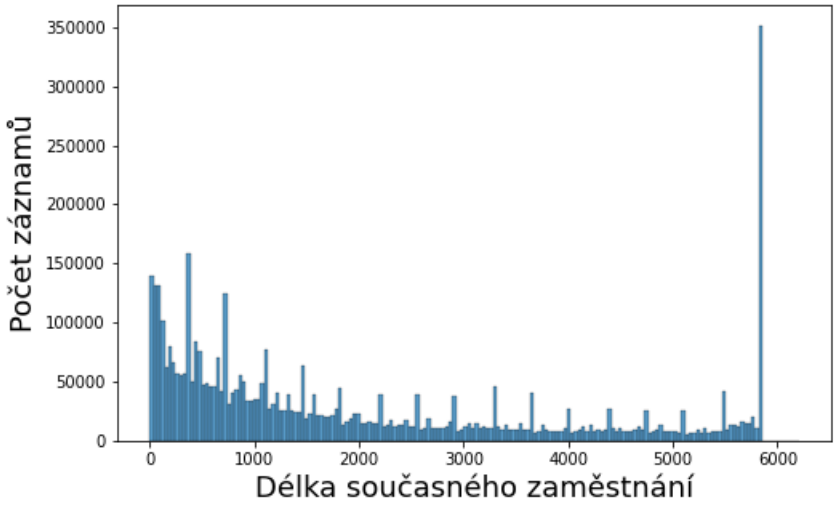

Figure 9.3: Histogram of frequency of entries (on the y axis) for various durations of current employment (on the x axis, in days).

Factor Impact

The factor is potentially significant, we recommend its implementation for consideration if data availability improves.

9.3.5 Duration of Inactivity

Data Availability

The duration of inactivity is available by calculations from database Extended STATMIN VZ. Inactivity is defined as an absence of an individual in database STATMIN VZ and at the same time in database STATMIN VZ OSVC (i.e., an individual who transitions from employment to self-employment and later returns to employment is not in inactivity). The analysis includes only persons who during their inactivity did not change their region nor their occupation, because a change in these factors reflects significantly on the value of earnings.

Factor Impact

The default hypothesis (implemented in the previous version of the model) predicted a decrease in assessment base in relation to the duration of inactivity or unemployment. Our analysis did not confirm this hypothesis.

The impact analysis of inactivity was divided into two parts depending on whether an employee changed their employer or not. Based on the expert judgement of the ordering party, these two cases differ typically especially in leaving the labour market. In case that an employer was changed (identified with employer identification number – column `ID_ORG_AN`), it is likely that the individual terminated their previous employment and joined new employment after some time, there would be no direct connection between the salary in the new employment and the salary in the previous employment. In contrast, a return to the same employer occurs most likely due to temporary interruption of employment for example for the reason of a maternity leave or a care for a dependent. In his case, the individual leaves the labour market but his position is “kept” by the employer who

counts with the employee returning to a similar position. Thus, salary after return from inactivity is directly linked to the previous salary.

Results of the analysis expressed in percentages of changes in income are portrayed in Figure 9.4. This analysis shows that the duration of inactivity has a certain impact on the assessment base and this impact differs based on whether there was a change in the employer or not. Specifically, after return to the same employer the salary increase is just under 5% for short inactivity duration, however, with longer inactivity this increase decreases and from three years of inactivity the change in salary is negative. In the second scenario (when the employer is changed after inactivity), the average new salary is approximately 2% higher for short inactivity, with longer inactivity this increase grows more and after three years the increase is 6% on average.

Although the total effect of inactivity and its duration on the assessment base is non-zero, it is relatively small. The biggest impact of a potential implementation lies especially in an increase in variability of assessment bases, not in any significant changes in the average assessment base. The factor also cannot be included in the first element of the wage equation (i.e., linear regression), because contrary to other factors, it is not its value which matters but it is its change (the event of return from inactivity). Adding this factor would not bring many advantages for the salary simulation, which is why the factor is not implemented.

The effect of a return from inactivity can be potentially better captured by another element that would be similar to a permanent shock (see below, section Permanent and Transitory Shocks), i.e., a one-time change in the assessment base with a permanent effect that persists also in further course of a given person's career. In such case, it would be suitable to distinguish the reasons for inactivity – whether the person was truly unemployed or for example on a maternity leave, and to select the value of this "shock" based on the reason.

We recommend considering an incorporation of the inactivity factor in the wage equation in the future.

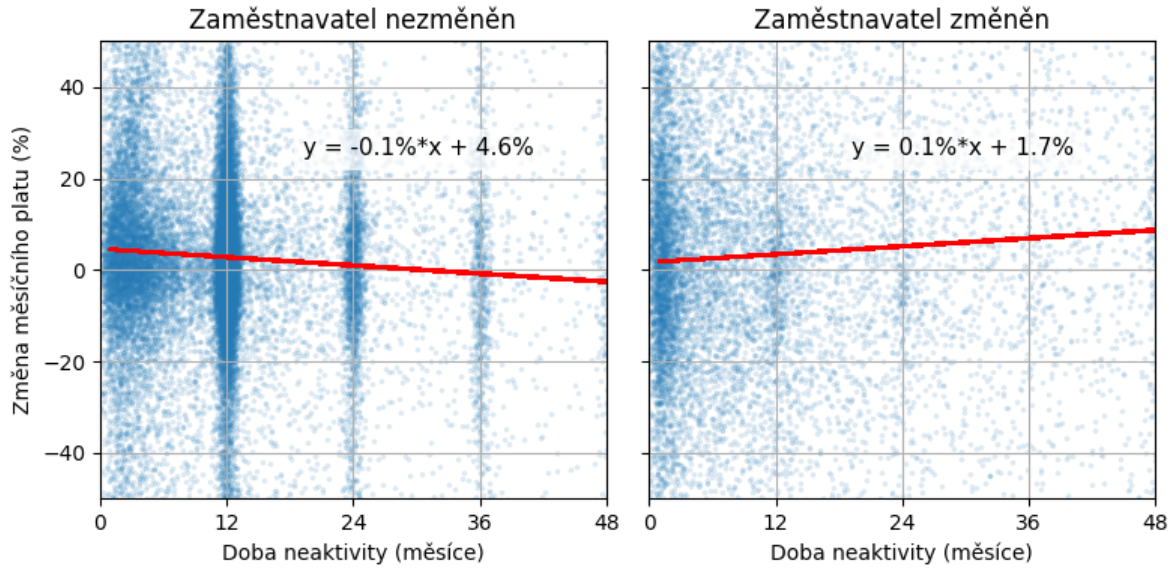

Figure 9.4: Change in monthly salary expressed in percentage, in relation to the duration of inactivity expressed in months. The left panel shows the analysed situation where an employer was not changed, the right panel shows the situation where an employer was not changed, the right panel shows the situation where an employer was changed. The inserted equation describes the straight line of linear

regression (red line) where y corresponds to the change in monthly assessment base in percentage and x corresponds to the duration of inactivity in months.

9.3.6 Education

Data Availability

Information about education is available in good quality in database Extended STATMIN VZ.

Factor Impact

The factor has a significant impact on the value of an assessment base since with higher level of achieved education results in increase of an average assessment base and an increase of its dispersion (see Figure 9.5).

This factor was implemented, thanks to its good availability.

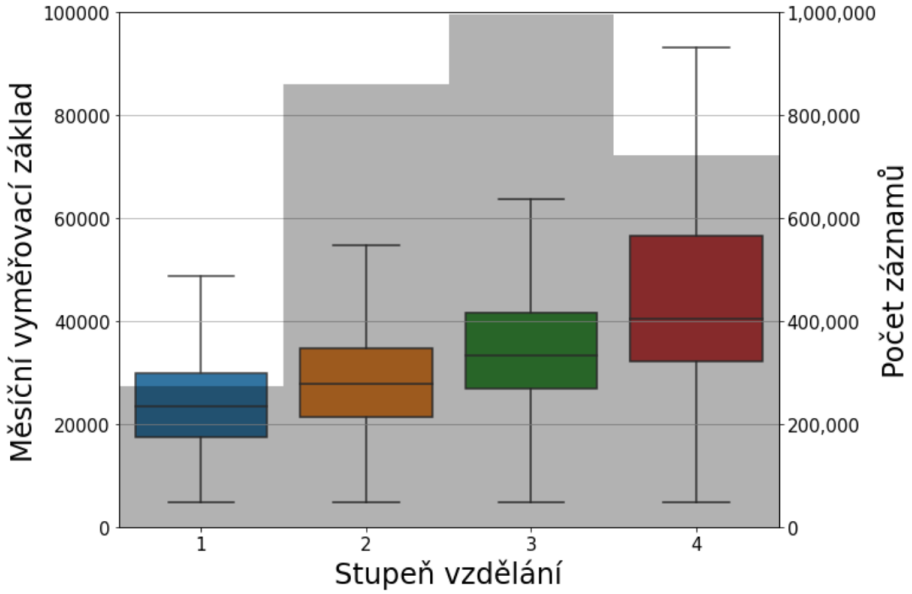

Figure 9.5: Education level 1 corresponds to primary education, education level 2 to secondary education without the completion of "maturita" graduation exams, education level 3 corresponds to secondary education with the completion of "maturita" graduation exams, and education level 4 represents university education. The illustration also shows a representation of the number of records in cleansed database Extended STATMIN VZ. Assessment bases are adjusted for the wage inflation coefficient.

9.3.7 Occupation

Data Availability

Information about occupation code is available in database Extended STATMIN VZ. We distinguish 40 types of occupations, each of them is defined by a two-digit code. These occupations can be classified into 9 groups – this classification is indicated by the first digit of the code.

Factor Impact

Based on the data analyses, the impact of the occupation code is crucial (Figure 9.6). There are differences in average monthly assessment bases and also their dispersions for individual occupations.

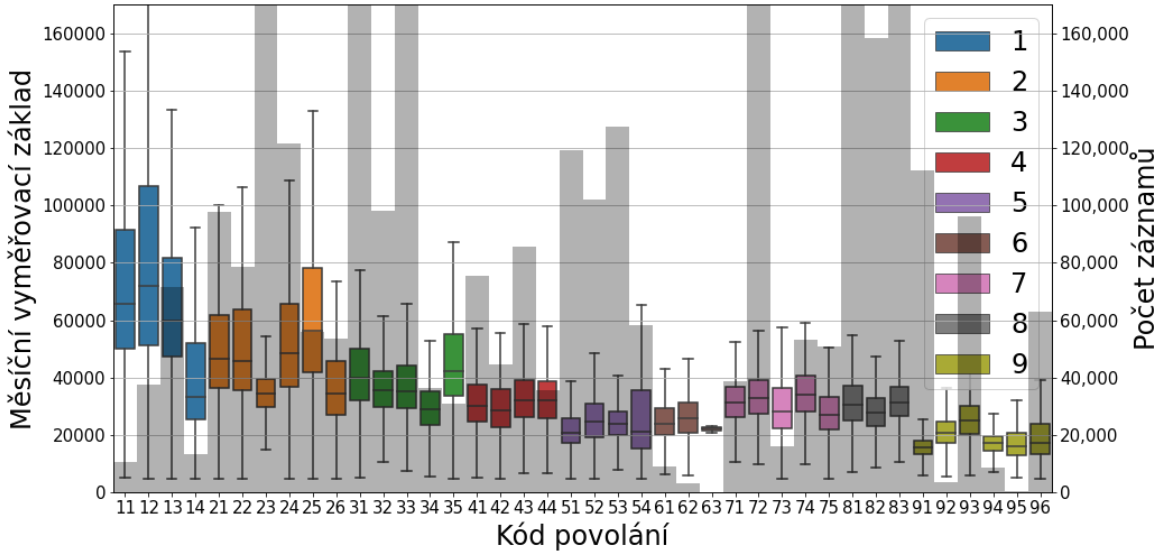

Figure 9.6: Analysis of monthly assessment bases in relation to occupation codes. The value of assessment bases significantly varies for different occupations. Occupations are classified into 9 groups which are illustrated by different colours. The illustration also shows the representation of the number of records in cleansed database Extended STATMIN. Assessment bases are adjusted for a wage inflation coefficient.

9.3.8 Region

Data Availability

The term region is used for a combination of a district and a locality (i.e., information whether the area is a rural or an urban area).

Information about postal code (PSČ) is available in database Extended STATMIN VZ. Information about region can be obtained by connecting a postal code codebook. In case of invalid postal codes, the relevant record cannot be used. The factor of region is available in excellent quality.

Factor Impact

The factor impact is medium, it does not play a fundamental role in the development of assessment base value (see Figure 9.7). The factor is implemented, its availability is very good.

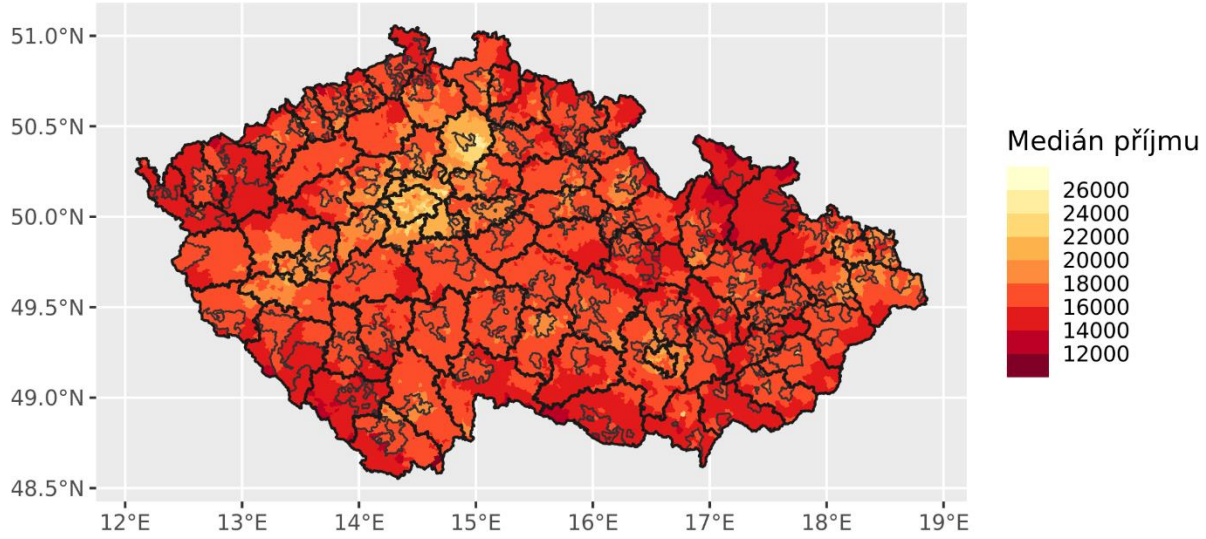

Figure 9.7: Analysis of monthly assessment bases in relation to geographical areas in the Czech Republic. The map shows a median of income.

9.3.9 Disability Pension Status

Data Availability

Database of paid out pensions STATMIN ANOD includes the same identifiers as database VZ, therefore it is possible to connect the databases. Based on database ANOD, we can determine whether each person was paid out an disability pension and the degree of disability. We also know the month when this pension was awarded, which means that we know in which months the given person was in disability pension. Thus, the availability of data is excellent.

Factor Impact

Paragraph §39 of the Act on Pension Insurance 155/1995 Coll. defines disability as a decrease in capacity for work of the insured person by at least 35%, individual degrees are then directly defined with an interval of the decrease. The first degree of disability pension corresponds to a decrease in capacity for work by 35 - 49%, the second degree corresponds to a decrease by 50 - 69% and the third degree to a decrease by at least 70%. This definition directly implies that a person in disability pension (depending on their degree of disability) shall have a significantly different income (on average lower) compared to a healthy person who does not differ in other factors (education, occupation, etc

The decrease in income for persons in disability pension is implemented as an additional coefficient which multiplies the salary predicted by the wage equation. This coefficient was derived as a ration of real monthly income (from database STATMIN VZ) and income predicted by the wage equation for a given working person who is in disability pension. Due to a numeric instability given by a low number of records, the values are smoothened by a moving average. The coefficient is calculated for each gender and age (from 18 to 65 years of age) separately, final values are in table `disability_salary.fac`, their graphical illustration is shown in Figure 9.8.

From the Figure 9.8, it is apparent that a difference between genders exists especially for younger people with disability of the third degree – there, the decrease in income for women is lower than for men. The degree of disability itself is the most significant factor, even though the decrease in income does not correspond to the legislative decrease in capacity for work. This difference is probably given by selection bias because the presented analyses does not at all reflect pensioners in disability who do not earn anything. Rather surprisingly, the income of invalid pensioners is with increasing age getting closer to the income of comparable healthy persons. This dependence was not further analysed; however, it is probably caused by increased selection bias – people with low income often terminate employment, or it could be influenced by decreasing income of healthy persons while the current income for an invalid pensioner is constant, or potentially by a combination of these causes.

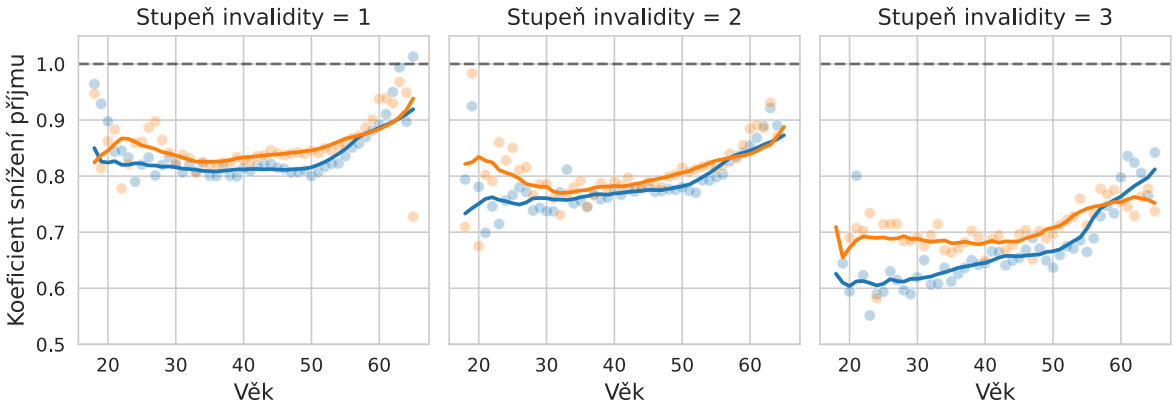

Figure 9.8: Analysis of decrease in income for working persons in disability pension, in relation to their age, gender and the degree of their disability. Each panel corresponds to one degree of disability. Individual values of decrease income coefficient are illustrated with points while final tabulated values adjusted for a moving average are outlined as solid lines. Blue colour corresponds to data for men, orange to data for women.

9.3.10 Student Status

Data Availability

Finding students who work during their studies is based on data from database Extended STATMIN VZ. The definition of a student in this context is a person whose current level of education (as reported by the employer) is lower than the level of education achieved later in life. The group of people who fall within this definition, however, includes a large amount of middle-aged people who are probably increasing their education and qualification but are not typical full-time students who would be working on the side. That is why, we limit the analysed group to individuals who have at least one record in the database stating their age below 25 years old. This limitation enables us to analyse also individuals whose employer report a change in education only retroactively after several years.

Factor Impact

We assume the possibility of students (especially university students) to work part-time during their studies. Considering the time demand of the studies, full-time work usually cannot be expected, i.e., students will not have a full-time work income such that we calculate from the wage equation. Thus, the predicated income is multiplied (similarly to people in disability pension) by a correction coefficient which is dependent on age, gender and achieved level of education of the person.

The results of the analysis are summarised in Figure 9.9, where the individual panels correspond to the highest levels of education. Points represent observed correction coefficients (real income divided by predicted income), solid lines are values smoothened by a moving average (with the use of 3 previous and 3 following values) and tabulated into table `work_school_salary.fac`.

The analysis shows an apparent trend of the students who are nearing the end of their studies are continuously (almost linearly) getting closer to the salary predicted by the wage equation. This trend is at most significant for university students (education code 4) who work part-time until the end of their studies around their 25 years of age. Students of secondary education institutions both with and without completing "maturita" graduation exams (codes 2 and 3) terminate their study around their 19 years of age. People with primary school as their highest achieved level of education do not yet have any opportunity to work during their studies and thus, using the coefficient for them does not make sense.

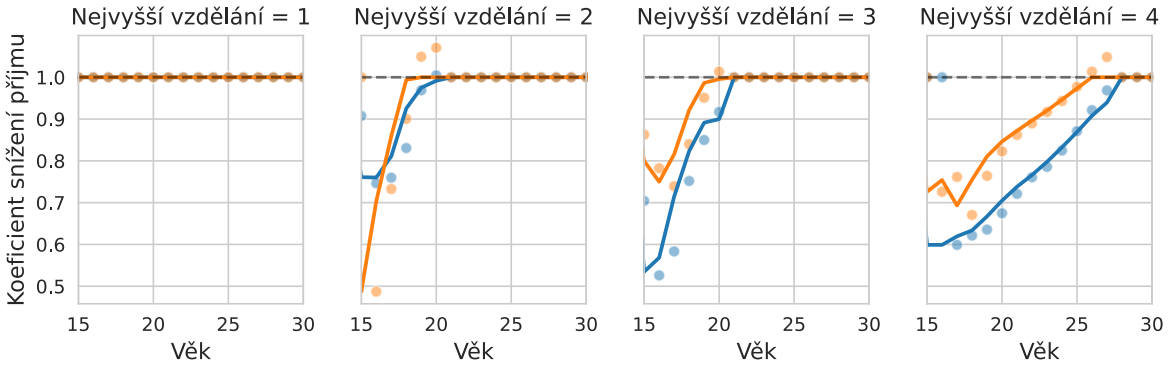

Figure 9.9: Analysis of students' income, in dependence on the highest achieved level of education, age and gender. Blue colour corresponds to data for men, orange colour to data for women. Individual values of the income decrease coefficient are illustrated with points, while final tabulated values adjusted for a moving average are shown as solid lines.

9.3.11 Care for a Dependent

Data Availability

Information whether a person cares for a dependent, alternatively whether they themselves are a dependent, is included in database PnP (see Chapter 'Care for Dependents'). This database does not have the same identifiers as database VZ and currently it is not possible to connect these databases. Therefore, it is not possible to decide whether a person in VZ cares for a dependent nor whether they are a dependent.

Factor Impact

Based on expert judgement and a discussion with the ordering party, we believe that care for a dependent has a negligible impact on work life of an individual. We assume that this impact is dependent primarily on the degree of dependence of the dependent. This impact probably influences mostly working hours and difficulty of performed work. We expect that these factors lead to considerably lower income for the caregiver.

A change in income of the caregiver was implemented similarly to implementations of changes for students and people in disability pension, in the form of a multiplication coefficient dependent on the

degree of dependent of the dependent. For the reason of absence of data, the coefficient was fixed at value 1 (i.e., no reduction of income is applied).

We recommend the factor to be analysed and potentially set to reduce the real value when the connection of databases VZ and PnP is possible.

9.4. Wage Equation

The creation of a wage equation was based on American CBOLT. This model introduces equations where on the left side from the equal sign, there is a natural logarithm of the modelled variable (*ln* E_{it}), which is gross monthly salary (or tax base for self-employed persons, no more further distinctions). Index *i* corresponds to a given person and index *t* to current time of the simulation. The right side of the equation consists of four elements

$$\ln E_{\mathrm{it}} = \ln \hat{E}_{it} + \mathrm{PED}_i + \sum_{s=1}^{t} \alpha_{is}\, \sigma_{\mathrm{perm}} + \beta_{it} \sigma_{\mathrm{trans}}$$

The first element represents a classical linear regression. Thus, this element models the average (logarithm of) gross salary in dependence on selected factors.

The second element, so called PED (*Permanent Earnings Differential*), expresses the average difference between logarithms of real and predicted gross salary in the last 5 years. Each person in the simulation (present from the start or new-born) gets assigned fixed values of `PED_EMPL` for the calculation of gross salary in employment and `PED_OSVC` for the calculation of tax base for self-employment.

The third and fourth elements of the wage equation are a permanent and transitory shock which enable variability of the income of a given individual across time. The permanent shock represents a change in individual abilities of the given person or a different random event which has an in-time cumulative effect (index *s* iterates though all previous steps of the simulation). The transitory shock represents random deviations of salary which cannot be otherwise explained by the model. Values α and β are random values determined in each step of the simulation for every person from standard normal distribution $N(0, 1)$, these values are subsequently multiplied by the values of standard deviations σ_{perm} and σ_{trans}, their calculations are described below.

9.4.1 Linear Regression

Linear regression was used to model salaries. Since salaries have log-normal distribution, we carried out a logarithmic transformation of the variable total real monthly assessment base. The logarithm of the assessment base already has a normal distribution (performed similarly as in model CBOLT) see Figure 9.10.

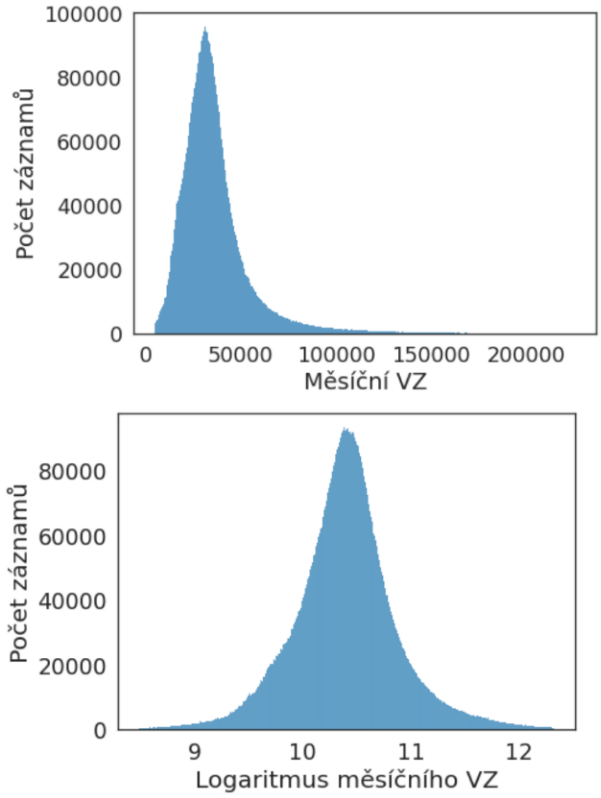

Figure 9.10: Distribution of a monthly assessment base and a distribution of the logarithm of a monthly assessment base.

A structure of the model had to be chosen, i.e., dependent variables and independent variables had to be selected. A dependent variable is a natural logarithm of the total real monthly assessment base which we want to model with the use of the independent variables. When it comes to independent variables, we analysed factors mentioned above. We explored multiple models with various selections of independent variables including interactions of age with education and age with occupation, also because similar interactions were used in the final report *Odvození parametrů rovnice mzdové dynamiky pro dynamický mikrosimulační model důchodového systému MPSV* (Trexima, 2015) *(in English: Derivation of parameters of wage dynamic equation for a dynamic micro-simulation model of the pension system of the Ministry of Labour and Social Affairs)*.

We selected the model with the best RMSE metric (root mean squared error) on „unseen" data, for which coefficients of regressions were not estimated (so-called test set, 20 % of data). Metric RMSE was the main criterium for the selection of the model.

The discovered model with the best RMSE metric used these independent variables:

- Age
- Second power of age
- Occupation (two-digit occupation code)

- Level of achieved education
- Combination of district and locality
- Interaction of age and occupation
- Interaction of age and education

Altogether, we obtained the following regression equation.

$$\begin{aligned}\ln\left(\hat{E}_{\mathrm{EMPL}}\right) = \ & \mathrm{INTERCEPT\_EMPL[sex]} \\ & + \mathrm{AGE} \cdot \mathrm{COEF\_AGE\_EMPL[sex]} \\ & + \mathrm{AGE}^2 \cdot \mathrm{COEF\_AGE\_SQUARED\_EMPL[sex]} \\ & + \mathrm{REGION\_EMPL[district, locality, sex]} \\ & + \mathrm{OCCUPATION\_EMPL[occupation, sex]} \\ & + \mathrm{AGE} \cdot \mathrm{OCCUPATION\_X\_AGE\_EMPL[occupation, sex]} \\ & + \mathrm{EDUCATION\_EMPL[education, sex]} \\ & + \mathrm{AGE} \cdot \mathrm{EDUCATION\_X\_AGE\_EMPL[education, sex]}\end{aligned}$$

The final model for self-employed persons was determined with a similar process. Contrary to database Extended STATMIN VZ, database STATMIN VZ OSVC does not include important predictors, such as occupation or achieved level of education. Therefore, we used these types of available independent variables:

- Age
- Second power of age
- Combination of district and locality

Similarly to employees, individual coefficient of linear regression were also estimated for men and women separately in case of self-employed persons. We obtained the following regression equation for self-employed persons.

$$\begin{aligned}\ln\left(\hat{E}_{\mathrm{OSVC}}\right) = \ & \mathrm{INTERCEPT\_OSVC[sex]} \\ & + \mathrm{AGE} \cdot \mathrm{COEF\_AGE\_OSVC[sex]} \\ & + \mathrm{AGE}^2 \cdot \mathrm{COEF\_AGE\_SQUARED\_OSVC[sex]} \\ & + \mathrm{REGION\_OSVC[district, locality, sex]}\end{aligned}$$

The final metrics R^2 and RMSE are displayed in the table below.

Table 7. Metrics of selected models for employees and for self-employed persons.

	Employees		**Self-Employed**	
	Men	Women	Men	Women
R^2	0,437	0,454	0,013	0,029
RMSE	0,356	0,338	0,798	0,761

When evaluating the predictive strength of the model for employees, we compared real observed data and results of predictions on CZK scale in dependence on gender, age and achieved level of education (see Figure 9.11) and in dependence on gender, age and occupation (see Figure 9.12). We can see that the selected regression can on average model assessment bases very accurately for various values of age, gender, education and occupation.

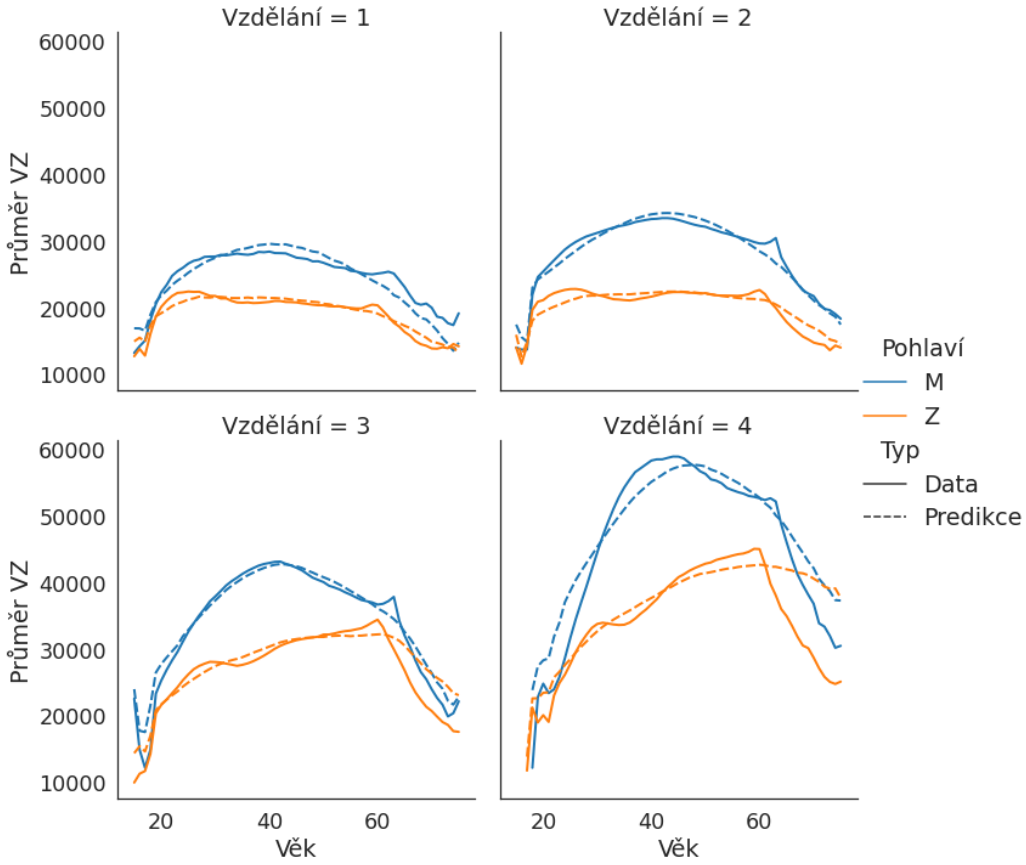

Figure 9.11: Comparison of average of real and predicted assessment bases of employees in dependence on gender (men in blue, women in orange), age and achieved level of education (1 primary, 2 secondary without "maturita" graduation exams, 3 secondary with "maturita" graduation exams, 3 secondary with "maturita" graduation exams, 4 university education) for observed data from database Extended STATMIN VZ.

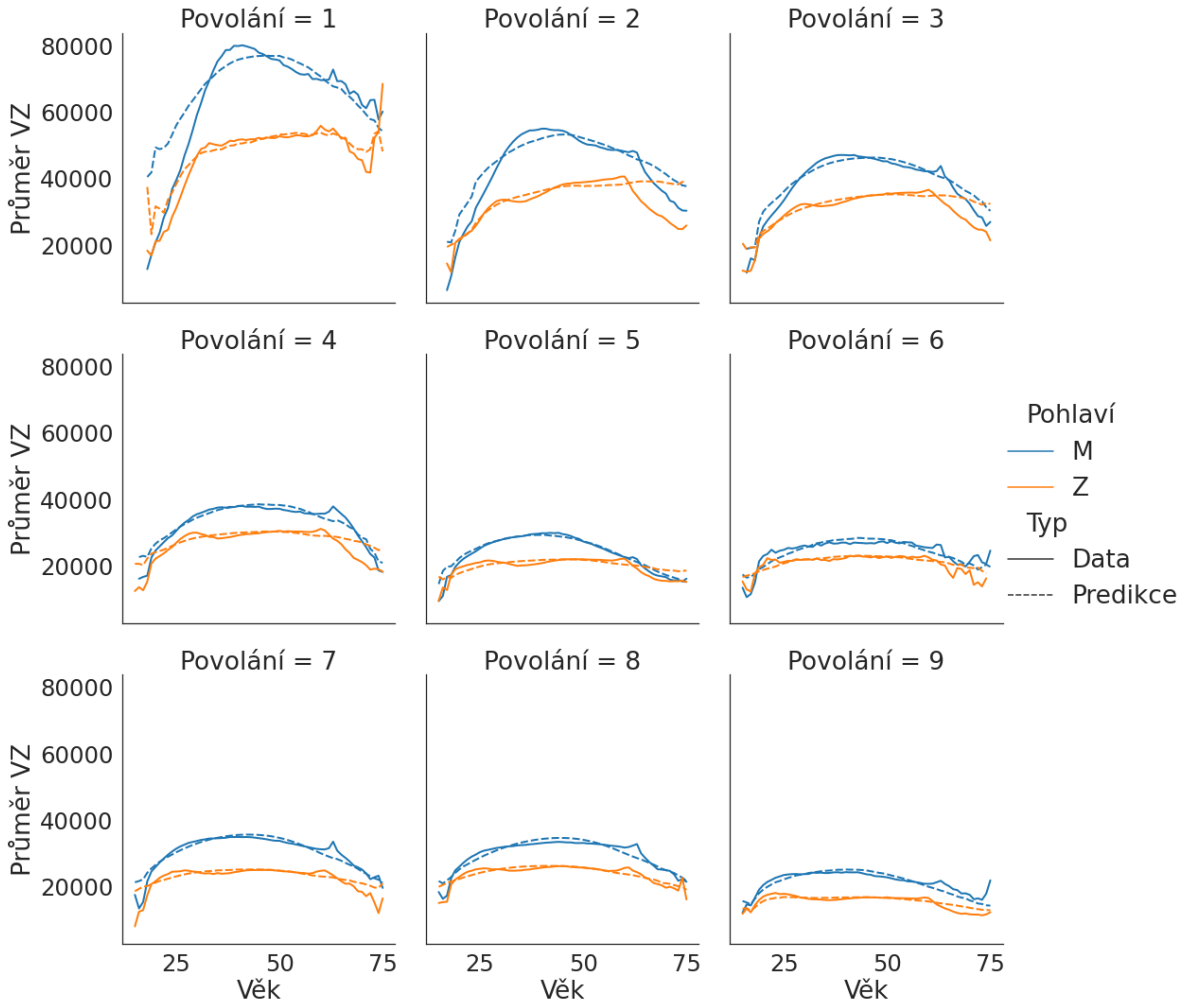

Figure 9.12: Comparison of averages of real and predicted assessment bases in dependence on gender, age and occupation code at granularity 1 (i.e., the first digit of the two-digit occupation code) from database Extended STATMIN VZ.

For self-employed, observed data and predictions were compared only in relation to age and sex because factors education and occupation are not available (see Figure 9.13).

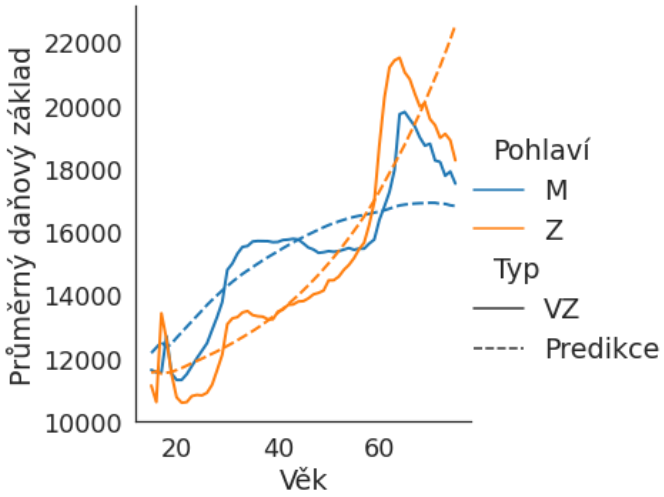

Figure 9.13: Comparison of averages of real and predicted tax bases of self-employed persons, in dependence on gender and age from database STATMIN VZ OSVC.

The following graphs show the size of residuals for both genders of employees (see Figure 9.14 and Figure 9.15). The residuals show differences between logarithms of real monthly salary and predicted monthly salary, displayed on logarithmic scale. There is no apparent structure (for example heteroscedasticity) visible in the residue graphs, therefore we can consider the selected model as suitable.

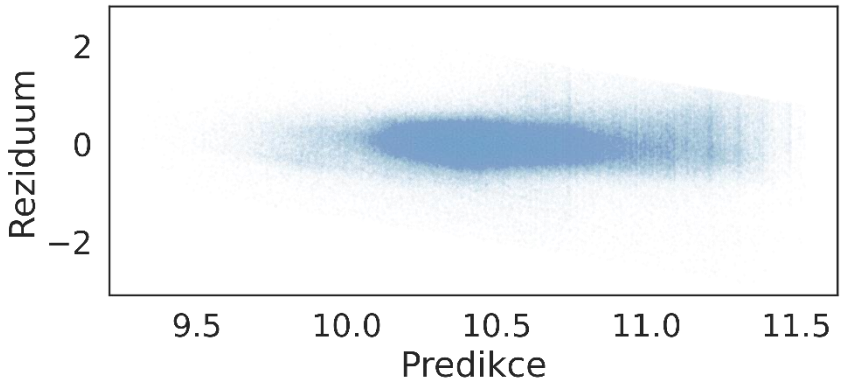

Figure 9.14: Comparison of predictions and residuals for employed men.

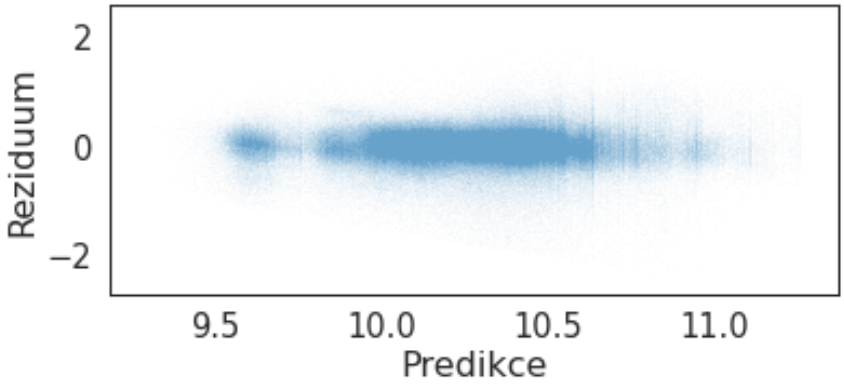

Figure 9.15: Comparison of predictions and residuals for employed women.

Residuals for self-employed persons are calculated in a similar way as for employees. They represent differences in logarithms of real tax base and predicted tax base on a logarithmic scale. Contrary to salary distribution of employees, the distribution of tax base for self-employed is has a lower boundary, and so residuals are also limited by this lower boundary (see Figure 9.16 and Figure 9.17).

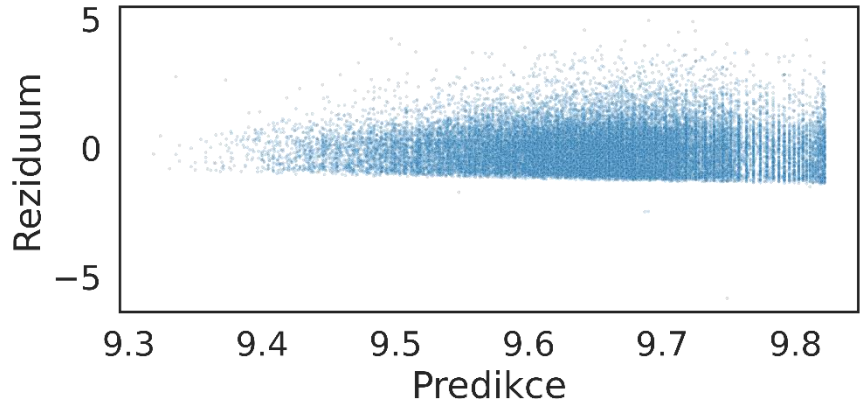

Figure 9.16: Comparison of predictions and residuals for self-employed men

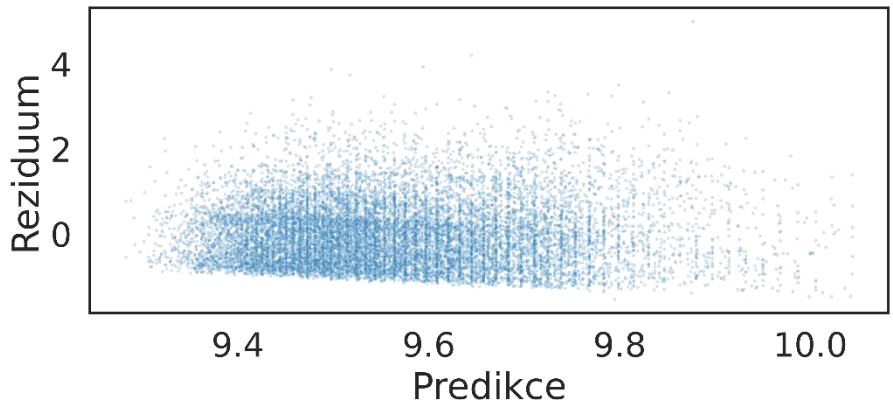

Figure 9.17: Comparison of predictions and residuals for self-employed women

9.4.2 PED: Permanent Earnings Differential

The value of PED is defined as an average difference between real income and its prediction in the last 5 years prior the first year of the simulation. PED plays a role as an individual effect and attempts to reflect individual variety in the model.

It is necessary to determine a value of `PED_EMPL` for each person in the model point database – this value states a constant individual contribution to the wage equation for an employee, a value of `PED_OSVC` must be determined, too, as it is contained in the calculation of tax base for self-employed persons. Both values are assigned to all model points, including those who never were employed (or never were self-employed) – as they might become self-employed or employed in the future. There are some persons for whom we model income even though we do not know the information necessary for direct calculation – in such cases, we determine the value PED in a different way. The method of assigning `PED_EMPL` and `PED_OSVC` is analogical, we will point out potential differences in the following text.

With respect to PED, we divide persons into the following categories, each of which uses a slightly different method of assigning PED.

1. Model points with records in database STATMIN VZ (alternatively STATMIN VZ OSVC for self-employed persons)
2. Model points without records in database STATMIN VZ (alternatively STATMIN VZ OSVC)
3. New-borns and immigrants

Individuals from the model point database who have records about their assessment (alternatively tax) base in the last five years, get the value of PED calculated based on these records using the following formula:

$$\mathrm{PED}_i = \frac{\sum_{\tau=1}^{N_i} \left(\ln E_{i,\tau} - \ln \hat{E}_{i,\tau}\right)}{N_i},$$

where *i* stands for a given individual, $E_{i,\tau}$ and $\hat{E}_{i,\tau}$ stand for a real and predicted assessment (tax) base in time τ and N_i is the total number of records in the last five years.

The equation describes the average difference between logarithms of monthly real and predicted assessment (tax) base in the last 5 years. For employees, these averages are calculated from data a month at a time. For self-employed, the averages are calculated a year at a time because self-employed persons have a maximum of one valid record about their assessment base for each year in STATMIN VZ OSVC, i.e., they have the same tax base in each month of a given year. A calculation performed with the use of individual months would give the same value of PED.

The following graphs show a comparison of logarithms of monthly assessment base and corresponding predictions before and after adding PED for employees (Figure 9.18) and for self-employed (Figure 9.19). Predictions for self-employed do not give as good results as predictions for employees, not even after adding PED. One of the reasons for that is the lack of useful factors for self-employed, especially occupation and education are missing, both being very important for the value of income for employees. Another reason is the conversion of original assessment bases to tax bases and the adjustments to their distribution (see above).

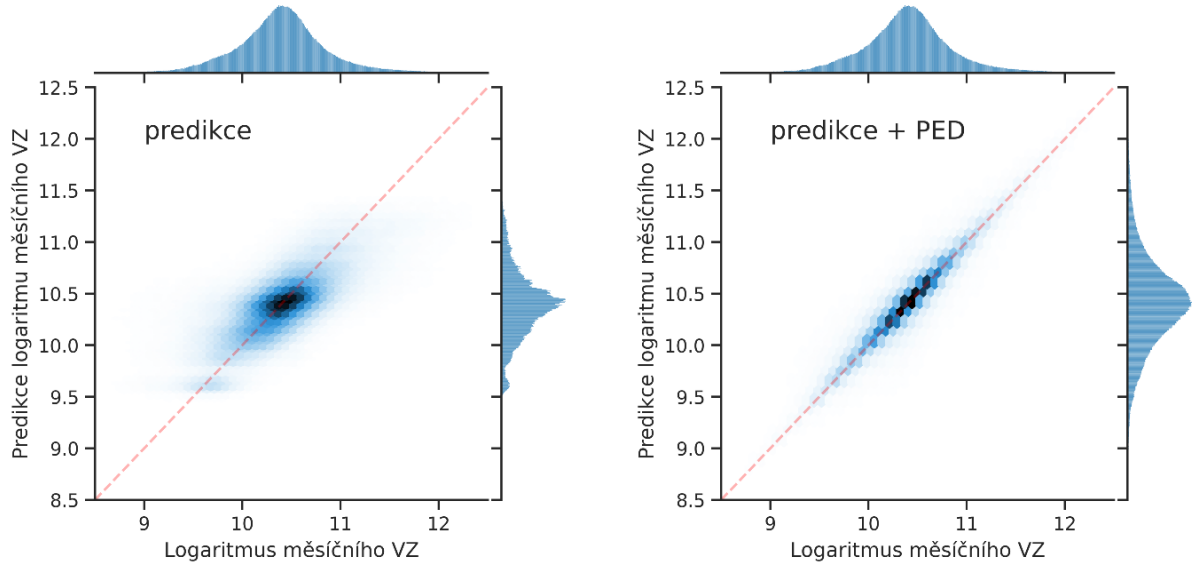

Figure 9.18: Comparison of real and predicted values of linear regressions before and after adding PED for employees

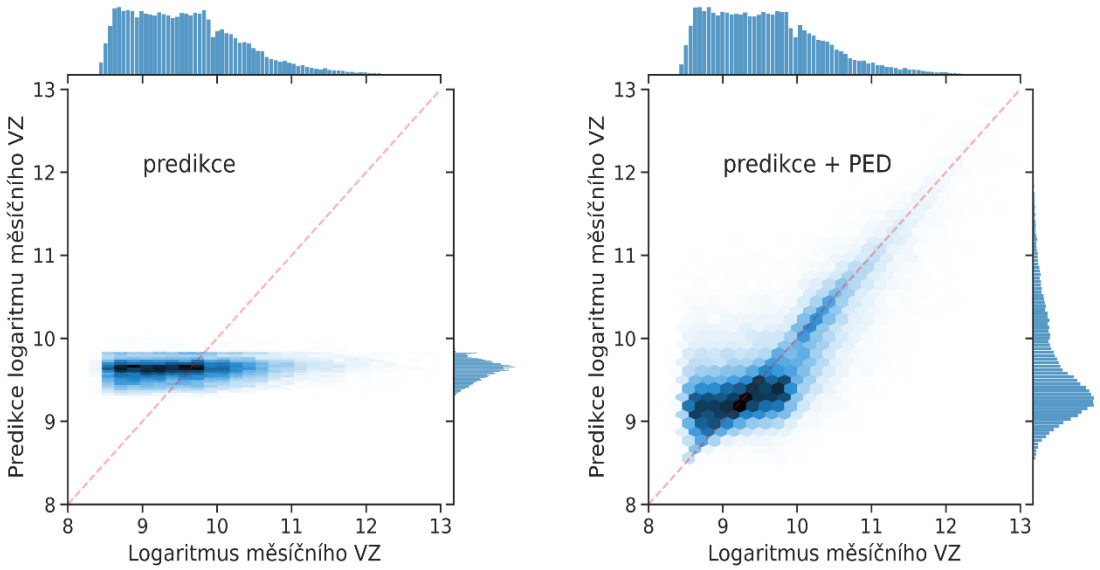

Figure 9.19: Comparison of real and predicted values of linear regressions before and after adding PED for self-employed persons.

PED must be determined differently for people who in the last five years do not have any records about their assessment base, or for those who enter the simulation as new-borns or immigrants. For them, PED is determined as a random number from a normal distribution with a zero middle value and with a standard deviation calculated based on distribution of PED obtained in the previous step.

The standard deviation for assigning PED was determined separately for individual categories of employees and self-employed persons. In case of employees, we determine this value for all combinations of gender and education. We do not know education for self-employed persons, and so

for them, we only distinguish gender. To calculate the standard deviation, we use only records from database Extended STATMIN VZ for which (in contrast to database STATMIN VZ) we know education and occupation with certainty.

The final standard deviations for every group are then saved in tables `PED_dist_empl.fac` (8 values) and `PED_dist_osvc.fac` (2 values).

9.4.3 Permanent and Transitory Shocks

The last two elements of the wage equation are permanent shock (σ^2_{perm}) and transitory (σ^2_{trans}) shock, which enable variability of a given person's income across time. A permanent shock represents a change in individual abilities of the person which change over time. A transitory shock expresses random deviations in salary which are otherwise not explained by the model.

During the simulation values of both shocks are determined in each step (i.e., each month) for every person as random numbers from normal distribution with a zero middle value and in-advance determined permanent (alternatively transitory) dispersion. Values of these dispersions are loaded up into tables and depend on factors gender, age (in categories 0-24, 25-34, 35-44 a 45+ years) and in case of employees also on education.

Values of the shocks are determined in the same way as in CBOLT model. The dispersion of difference in salary between two moments distanced from each other by d months depends linearly on the values of both shocks – the permanent shock gives a direction of the straight line and transitory shock is half the intersection of the straight line with the y axis.

$$\text{var}(\ln \text{E}_{\text{t+d}} - \ln \text{E}_{\text{t}}) = \text{d}\sigma^2_{\text{perm}} + 2\sigma^2_{\text{trans}}$$

Calculations were carried out separately for employees and for self-employed persons. The range of value d was limited with a lower boundary of 12 months (many employees' salaries change in yearly rhythm and dispersions under 12 months are very unstable) and with an upper boundary of 84 months (i.e., the length of database Extended STATMIN VZ).

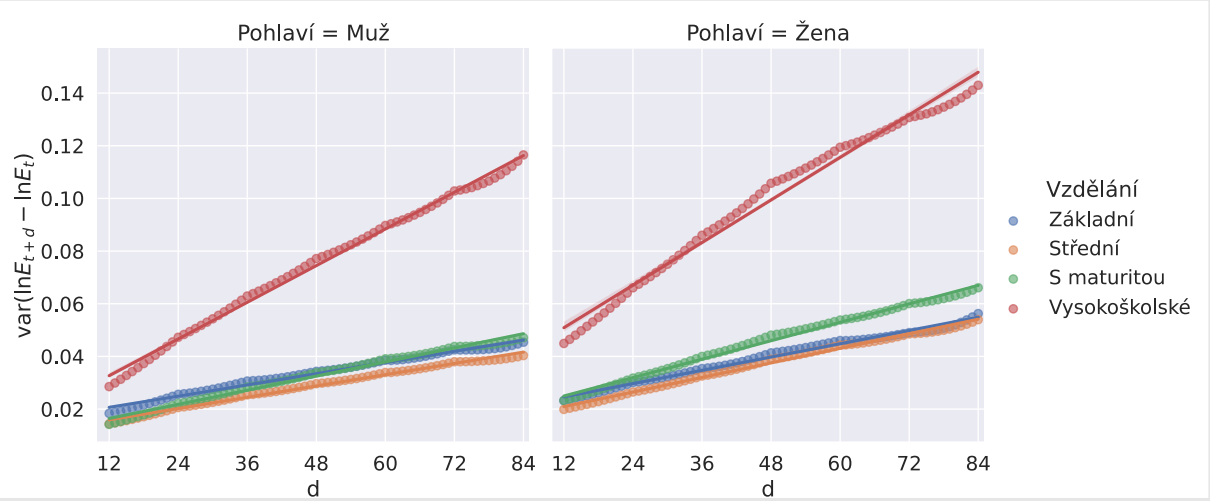

Figure 9.20: A representative example of deriving of permanent shock values and transitory shock values for employees. The permanent shock corresponds to the incline of the straight line, the transitory shock corresponds to a half of the intersection with y axis. Real shocks are also dependent on age.

9.5. Implemented Changes

9.5.1 Preparation of Model Points

Two new columns were added into the model point database: `PED_EMPL` for a calculation of gross salary of an employed person, and `PED_OSVC` for a calculation of tax base of a self-employed person. Each person gets both of these values determined. This section describes the process of the calculations for all persons in model points.

Employees

Data from database STATMIN VZ were used for preparation of `PED_EMPL`. Also, data from the model point database (MP) were used including information about region and occupation newly added as per chapters 5 (Region) and 6 (Occupation) delivered in this order.

First, we calculate PED for people who in the last five years have records in STATMIN VZ. Then, we add value PED, stochastically based on known PED distribution for individual genders and levels of education, to other persons from the model point database.

Data from database STATMIN VZ are processed in a similar way as were data from database Extended STATMIN VZ with the exception of occupation and education which were not included in data of STATMIN VZ.

Specifically, we selected the following records from database STATMIN VZ:

- In years 2015 until 2019, including.
- With an entered non-zero assessment base.
- From employment, i.e., first digits of gainful activity code (KVČ) in range 1-9, because income is modelled as salary from the main employment.
- For some (at least in database MP), information about postal code (PSČ), occupation and education exists. If a record does not have an entered value of occupation or postal code, this will be first added from a different record of the same person from the nearest preceding to the one in question, alternatively the nearest one that follows.

A total number of days in employment is then calculated with the deduction of excluded time period from the difference between DO (until) and OD (from), and a daily assessment base is calculated with a division of the assessment base by the total number of days.

Data are divided into months and then aggregated in such a way that for each person and month contained in data, there is a unique record which includes information about the total daily assessment base, gender, year of birth and postal code. Assessment bases are converted from nominal to real to year 2019 with wage inflation coefficients (coefficients of increase in general assessment base valid for pensions which were awarded in year 2019), and then these daily assessment bases are converted to monthly assessment bases.

Information about education and occupation is obtained from MP, in case that a postal code is unavailable in database STATMIN VZ, it is added also from MP. We assume that each person monthly earns at least 10 000 CZK, which is why the value of assessment base is limited with a lower boundary of this value (records with lower total monthly assessment base are removed), so that PED is not distorted in case that a given individual had a part-time employment with low earnings.

To the prepared data, a model of linear regression, with previously obtained coefficients, was applied, which gave the value of a logarithm (and then the real value) of the assessment base. This predicted logarithm of income was used then used for calculating PED (see above).

This process calculated PED for 61.3 % people from MP. The remaining 38.7% people from MP had PED selected randomly from a prepared distribution based on their gender and education (see above).

9.5.2 Self-Employment

The calculation of `PED_OSVC` comes from an adjusted database STATMIN VZ OSVC (see section Data Sources) with yearly granularity.

The prepared linear-regression model is used for each ID and year (in interval 2015 - 2019) and a predicted logarithm of a monthly tax base is obtained. We compare it with the logarithm of real tax base (i.e., double the monthly assessment base, obtained by dividing a yearly assessment base by the number of months during which a given person was working and adjusted for inflation to year 2019). PED is calculated as an average value of the difference of real and predicted monthly tax base logarithms in the last 5 years. This process assigns the value of `PED_OSVC` to 9.7 % persons from MP.

The value of PED for model points which in the last five years did not have any income from self-employment is assigned stochastically from a normal distribution with zero average and dispersion calculated from the distribution of PED calculated in the previous step (see Figure 9.21). A deviation of the PED distribution to the left (compared to a normal distribution) is mostly an artefact that comes from data preparation, when persons who showed a yearly assessment base at the level of a minimum assessment base are assigned a lower net income and this income is also limited by a lower boundary (see section Preparation and Database Cleansing: Self-Employed).

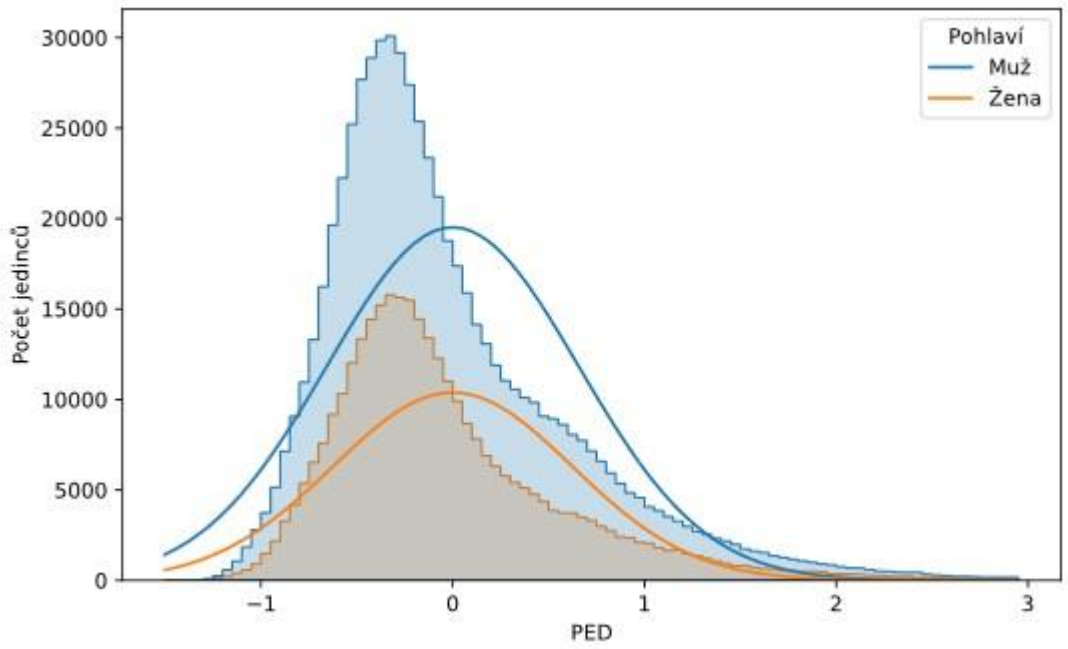

Figure 9.21: Distribution of PED for self-employed persons. The displayed areas correspond to histograms of PED calculated from STATMIN VZ OSVC. Lines represent distribution which were used for persons who did not have a history of income from self-employment.

9.5.3 DCS

In respect to salaries, it is necessary to extend model points for Prophet with columns `PED_EMPL` and `PED_OSVC`. Similarly to previous chapters, database `INEP_PARTICIPANTS` is extended with DCS `MERGE_INEP_EXT_INPUTS` for existing persons. There are no changes carried out in other DCS in relation to already existing persons.

In case of newly generated persons (children and immigrants), it was necessary to calculate values for newly added columns `PED_EMPL` and `PED_OSVC`. The variable `PED_EMPL` is calculated based on a person's gender and achieved level of education from values in table `PED_dist_empl.fac`. Similarly, `PED_OSVC` is calculated from table `PED_dist_osvc.fac` (only based on gender). Vales obtained from the `mentioned.fac` tables are multiplied by a random number from a normal distribution. This calculation is performed in DCS `03_newborn_and_children.DCS` and `04_immigrants.DCS`.

9.5.4 Description of .fac Tables

Disability_salary.fac

This table was only updated, it includes coefficients for wage reduction in case of disability, it is a shared table for both employees and self-employed persons.

Table 9.1: Structure of Table Disability_salary.fac

Code	Comment
SEX	Gender
AGE_NOW_Y	Current age
DISABILITY_LEVEL	Degree of disability
DIS_SAL_RATIO	Coefficient for salaries of people in disability pension

Education_age_empl.fac

A table with coefficients `EDUCATION_X_AGE_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.2: Structure of Table Education_age_empl.fac

Code	Comment
EDUCATION	Education
SEX	Gender
COEFFICIENT	Value of coefficient

Education_empl.fac

A table with coefficients `EDUCATION_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.3: Structure of Table Education_empl.fac

Code	Comment
EDUCATION	Education
SEX	Gender
COEFFICIENT	Value of coefficient

Occupation_age_empl.fac

A table with coefficients `COEF_OCCUPATION_X_AGE_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.4: Structure of Table Occupation_age_empl.fac

Code	Comment
OCCUPATION	Occupation
SEX	Gender
COEFFICIENT	Value of coefficient

Occupation_empl.fac

A table with coefficients `OCCUPATION_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.5: Structure of Table Occupation_empl.fac

Code	Comment
OCCUPATION	Occupation
SEX	Gender
COEFFICIENT	Value of coefficient

General_empl.fac

This table contains coefficients `INTERCEPT_EMPL`, `COEF_AGE_EMPL` a `COEF_AGE_SQUARED_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.6: Structure of Table General_empl.fac

Code	Comment
COEF_NAME	Name of coefficient
SEX	Gender
COEF_VALUE	Value of coefficient

General_osvc.fac

This table contains coefficients `INTERCEPT_OSVC`, `COEF_AGE_OSVC` a `COEF_AGE_SQUARED_OSVC` for the calculation of the first element of the wage equation for self-employed persons (see above).

Table 9.7: Structure of Table General_osvc.fac

Code	Comment
COEF_NAME	Name of coefficient
SEX	Gender
COEF_VALUE	Value of coefficient

Region_empl.fac

This table contains coefficients `REGION_EMPL` for the calculation of the first element of the wage equation for employees (see above).

Table 9.8: Structure of Table Region_empl.fac

Code	Comment
REGION	Code of region
LOCALITY	Code of locality
SEX	Gender
COEFFICIENT	Value of coefficient

Region_osvc.fac

This table contains coefficients `REGION_OSVC` for the calculation of the first element of the wage equation for self-employed persons (see above).

Table 9.9: Structure of Table Region_osvc.fac

Code	Comment
REGION	Code of region
LOCALITY	Code of locality
SEX	Gender
COEFFICIENT	Value of coefficient

Shock_age.fac

This table is used for age categorisation for the following tables with permanent and transitory shocks.

Table 9.10: Structure of Table Shock_age.fac

Code	Comment
CATEGORY	Code of age category
MIN_AGE	Lower boundary of age category

Shock_perm_empl.fac

Table with coefficients sigma for the calculation of a permanent shock in the wage equation for employees (see above).

Table 9.11: Structure of Table Shock_perm_empl.fac

Code	Comment
AGE_NOW_Y	Current age
SEX	Gender
EDUCATION	Education
PERM_SIGMA	Coefficient sigma

Shock_trans_empl.fac

Table with coefficients sigma for the calculation of a transitory shock in the wage equation for employees (see above).

Table 9.12: Structure of Table Shock_trans_empl.fac

Code	Comment
AGE_NOW_Y	Current age
SEX	Gender
EDUCATION	Education
TRANS_SIGMA	Coefficient sigma

Shock_perm_osvc.fac

Table with coefficients sigma for the calculation of a permanent shock in the wage equation for self-employed persons (see above).

Table 9.13: Structure of Table Shock_perm_osvc.fac

Code	Comment
AGE_NOW_Y	Current age
SEX	Gender
EDUCATION	Education
PERM_SIGMA	Coefficient sigma

Shock_trans_osvc.fac

Table with coefficients sigma for the calculation of a transitory shock in the wage equation for self-employed persons (see above).

Table 9.14: Structure of Table Shock_trans_osvc.fac

Code	Comment
AGE_NOW_Y	Current age

SEX	Gender
EDUCATION	Education
TRANS_SIGMA	Coefficient sigma

Work_school_salary.fac

This table contains coefficients of wage reduction for working students. The table extended from age zero to `MAX_TABLE_AGE`, because the termination of study is given in MP (`EDUCATION_FINISH_AGE`) and it can be arbitrary. `WORK_SCHOOL_SAL_RATIO` is equal to zero until 14 years of age, `WORK_SCHOOL_SAL_RATIO` is equal to one for other missing values. This table is a shared table for employees and self-employed persons.

Table 9.15: Structure of Table Work_school_salary.fac

Code	Comment
SEX	Gender
AGE_NOW_Y	Current age
EDUCATION_MAX	Highest achievable level of education
WORK_SCHOOL_SAL_RATIO	Coefficient for calculation of wage

PED_dist_empl.fac

This table includes the value for the calculation of PED for employees based on gender and achieved level of education.

Table 9.16: Structure of Table PED_dist_empl.fac

Code	Comment
EDUCATION	Achieved level of education
PED_MALES	Value of PED for men and a given level of education
PED_FEMALES	Value of PED for women and a given level of education

PED_dist_osvc.fac

This table includes the value for the calculation of PED for self-employed persons based on gender.

Table 9.17: Structure of Table PED_dist_osvc.fac

Code	Comment
SEX	Gender of a given person
PED_STD	Value of PED for a given person

9.5.5 Prophet

A wage equation was implemented into the model according to the analysis of DataSentics (described above), separately for employees and self-employed. A permanent and a transitory shock is calculated every month, random numbers can be either modelled with the use of a set generator of random

numbers (the same as for example in `STOCH_EVENTS`) or newly generated each time. The permanent shock is not adjusted to the current value if a given person is not in an employed status. The wage calculated in the model is increased by a wage inflation.

Moreover, variables for gross salary of an employee (alternatively a tax base for a self-employed) – `GRS_SAL_EMPL` (alternatively `GRS_SAL_OSVC`) were adjusted in Prophet. For employees, there is a limitation by a lower boundary to minimum wage, there is no limitation applied for self-employed persons (self-employed get a minimum applied to the assessment base). The model enables a reduction of the calculating values – both for people in disability pension in a similar manner, newly by this process also for students (details are in the analysis above) and also for caregivers who care for a dependent (currently set without a reduction). The same reduction percentage is used for both employees and self-employed.

9.6. Implementation Feasibility Assessment of Additional Factors

The implementation of additional factors can be split into two categories. The first is an implementation of changes which only expand the wage equation without any fundamental changes to the model needed. All non-implemented factors named in the table below with the exception of self-employment belong to this category. The implementation of these factors would entail primarily an addition of a new table such as `Education_empl.fac`. It would be also necessary to add "a categorisation table" for numerical variables such as `Shock_age.fac`. The second category of changes is the implementation of the factor of self-employment which, too, would expand the wage equation but for which it is also expected a parallel implementation of a change in self-employment, which requires a more significant intervention to the model. This change in the model is possibly at a similar level of difficulty as the implementation of an occupation change (Chapter 6).

9.7. Summary and Evaluation of Other Factors

9.7.1 Limitations of the Current Approach

We consider it important to mention data limitations of the used approach. Coefficients of linear regressions for employees were estimated based on data from database Extended STATMIN VZ. This database, however, has a different distribution of salaries than the larger database STATMIN VZ (see Figure 9.22. Predicted salaries, therefore, might be slightly

over-valued. Obtaining information about education and occupation of all persons in database STATMIN VZ could build a model which would be more representative for the entire population.

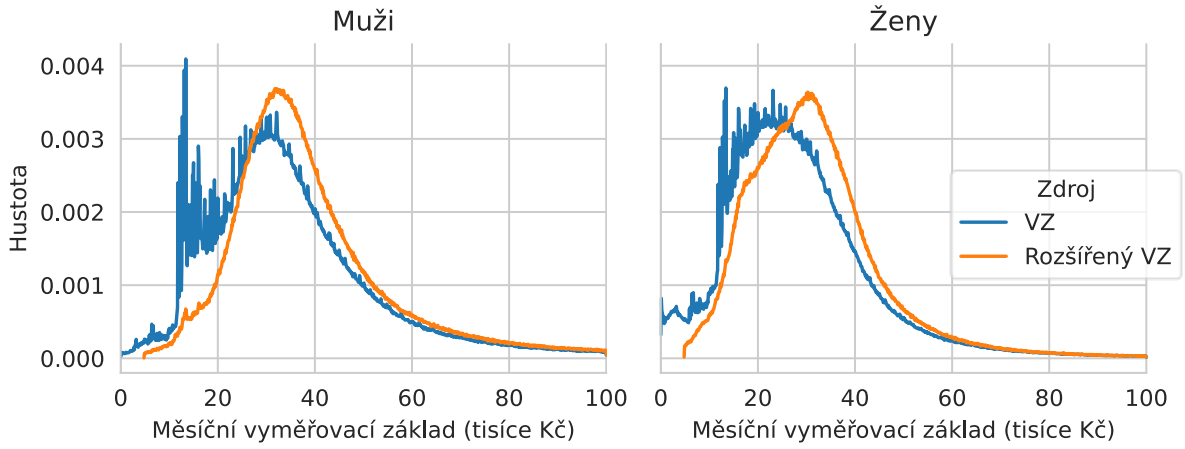

Figure 9.22: Comparison of distributions of assessment base in STATMIN VZ (blue curve) and in database Extended STATMIN VZ (orange curve).

Also, information about full-time / part-time employment of persons in the databases would help to obtain a more accurate model – the current approached assumes full-time employment for all persons, which is not completely realistic.

The model has only a low predictability strength for self-employed individuals, and thus any new piece of information which could be incorporated into STATMIN VZ OSVC should be verified for its ability to increase accuracy of this model. This can be for example a branch of self-employment or an achieved level of education.

Information (factors) included in the model and factors recommended for additional implementation are summarised below.

9.7.2 Employees

Table 9.18: Suitability of future implementation of additional factors - Employees

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Gender	Excellent	High	Implemented	Implemented
Age	Excellent	High	Implemented	Implemented
Occupation	Excellent	Very high	Implemented	Implemented
Education	Excellent	High	Implemented	Implemented
Region	Excellent	Medium	Implemented	Implemented

Total Duration of Employment	Medium	High	Low	We recommend comparing it with the significance of the age variable
Duration of Current Employment	Medium	High	Low	We recommend comparing it with the significance of the age variable
Full-Time / Part-Time Employment	Not available	Medium	Low if available	If data becomes available, we recommend for consideration
Employment Inactivity	Excellent	Medium	Medium	We recommend considering its implementation in the permanent shock
Disability Pension Status	Excellent	High	Implemented	Implemented
Student Status	Medium	Medium	Implemented	Implemented
Care for a Dependent	ID cannot be assigned	Unknown	Implemented	We recommend for an analysis

9.7.3 Self-Employed Persons

Table 9.19: Suitability of future implementation of additional factors - Self-Employed Persons

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Gender	Excellent	High	Implemented	Implemented
Age and its Second Power	Excellent	High	Implemented	Implemented
Branch of Self-Employment	Not available	Unknown	If data available - low for implementation in the wage equation, high for current implementation of occupation change	We recommend considering this factor if data becomes available
Implemented	Not available	Unknown	Low if available	We recommend considering this factor if data becomes available

Region	Excellent	Medium	Implemented	Implemented
Duration of Self-Employment	Medium	Unknown	Low	We recommend comparing it with the significance of the age variable
Self-Employment Inactivity	Excellent	Medium	Low	We recommend considering its implementation in the permanent shock
Disability Pension Status	Excellent	High	Implemented	Implemented
Student Status	Low	Medium	Implemented	Implemented
Care for a Dependent	ID cannot be assigned	Unknown	Implemented	Implemented

10. Chapter 8, Self-employment

10.1. Introduction

Main aims of Chapter Self-Employment were to enable a combinatory activity of self-employment and employment, to categorise self-employed persons into a subgroup of self-employment as a primary activity or a subgroup of self-employment as a secondary activity and lastly, to update probabilities of current processes. The outcomes are an analysis of the possible self-employment and employment combination, a calculation of probability that a person becomes self-employed, and an assessment of factors influencing self-employment and employment. Furthermore, information about self-employment was added to a model point database. The model was extended to include the possible combination of self-employment and employment gainful activities, also conditions for self-employment as a secondary activity were added, including a referral to the reason for being in self-employment as a secondary activity.

10.2. Source Data

Data, based on which the analyses and calculations were carried out, come from databases VZ STATMIN and VZ STATMIN OSVČ. During secondary analyses, data was compared with results from The Czech Social Security Administration (ČSSZ) or the Ministry of Industry and Trade (Ministerstvo průmyslu a obchodu), specifically in calculation of a probability that an immigrant becomes self-employed. Moreover, comparisons with data from the Czech Statistical Office (Český statistický úřad) were used for probabilities that a new-born becomes an employee or self-employed during their lifetime – public database, Employment According to Occupation.

INEP database was used for secondary analyses in chapter Self-Employment. For the probability calculations as such, only VZ STATMIN OSVČ database was used - VZ STATMIN OSVČ database includes more detailed information about individual self-employed persons and enables observations of the activity development, also in relation to employment (VZ STATMIN database), with monthly granularity (enables to observe whether a gainful activity was performed for example only for 6 months, 4 months, and the like). INEP database provides information about duration of such an activity with a yearly time granularity (it automatically assumes that the activity was actively performed the whole year, even in case of only performing the activity for a fraction of the year). In order to obtain data with higher level of detail and a sufficient time period for analysis (since year 2013), VZ STATMIN OSVČ database was used.

10.3. Basic Conditions for Self-Employment in the Czech Republic

Within the Self-Employment chapter, conditions of commencing self-employment as set out by the Czech legislation were analysed, so were forms of self-employment (primary or secondary gainful activity) and a combination of self-employment with employment. Analyses of available data content were performed together with a detection of what self-employment parameters are available in the data. These analyses gave the following results.

10.3.1 Basic Definition of Self-Employment, Differentiation and Declaration of Income

Czech legislation defines self-employment performed as either a primary gainful activity or as a secondary gainful activity. In order to engage in self-employment, it is necessary to first obtain a relevant permit as set out by the Czech legislation. It is not possible to register as self-employed for

those who have not completed compulsory schooling. There is also a minimum age requirement which is set at reaching 15 years of age.

In order to register self-employment as a secondary gainful activity, conditions for self-employment as a secondary gainful activity must be met. Fulfilment of those conditions, however, does not give an automatic entitlement for its registration.

10.3.2 Significant Characteristics of Self-Employment for the Purposes of Further Analyses

A person registered to self-employment as their primary gainful activity is obligated to pay monthly advance pension insurance payments. This fact proved to be significant for the purposes of the simulation. The minimum payment which is required depends on their average monthly income. Registering for sickness insurance, under The Czech Social Security Administration, is voluntary.

A person registered to self-employment as their secondary gainful activity is not obligated to pay advance pension insurance payments unless their minimum assessment base exceeds the limit stated by relevant legislation (the value changes based on average salary which gets yearly recalculated at a national level).

If the assessment base of a self-employed person, for whom self-employment is a secondary gainful activity, does not exceed the above-mentioned limit, they can register to voluntary pension insurance. A self-employed person, for whom self-employment is a secondary gainful activity can also register to voluntary sickness insurance.

10.3.3 Conditions for Registration to Self-Employment as a Secondary Gainful Activity

Self-Employment is considered (as of 2021) (ČSÚ, 2021) as secondary gainful activity if the following holds for the self-employed person in the relevant calendar year:

- They were employed,
- They were entitled to disability pension payments or to old-age pension,
- They were entitled to parental allowance or to maternity benefits or to sickness benefits due to pregnancy and birth giving, that is if these benefits came from employee sickness insurance,
- They were a primary care giver of a person younger than 10 years of age who is dependent on another person's care at first degree level (light dependency)
- They were a primary care giver of a person who is dependent on another person's care at second, third or fourth degree if the person is a close contact or the person is not a close contact but lives in the same household as the self-employed person,
- They performed military services in the Czech Republic Armed Forces if not soldiers by occupation or civil service,
- They themselves were a child dependent.

Conditions for self-employment as a secondary gainful activity did not fundamentally change in the analysed period from year 2013 until 2019.

10.3.4 Self-Employment Assessment Base and Minimum Assessment Base for Advance Payments towards Pension Insurance

Statement of income and expenses from self-employment used definitions of calculated assessment base, partial assessment base, minimum assessment base and determined assessment base. For the purposes of the analysis, it was found out that available data in VZ STATMIN OSVČ database include determined assessment base.

Determined assessment base is defined as an amount which is referred to in calculation of future pension. In this case, it mostly corresponds with the minimum assessment base or it is an amount set by the self-employed person themselves – if it is an amount set by the self-employed person, it must be higher than the minimum assessment base and at the same time it must be equal to or higher than the calculated assessment base.

The minimum assessment base is a product of the number of months in which self-employment as a primary activity is performed and a minimum monthly assessment base for self-employment as a primary activity. In case of self-employment as a secondary activity, determination of the minimum assessment base corresponds to the determination described for self-employment as a primary activity. The minimum assessment base is determined for each (primary and secondary activity) separately as a determined percentage of average salary for the given year at a national level).

Minimum Yearly Assessment Base for Calculation of Advance Payments towards Pension Insurance

Table 10.1: Minimum assessment base in self-employment as either a primary or secondary gainful activity

Year	Primary Gainful Activity	Secondary Gainful Activity
2013	77 652	31 068
2014	77 832	31 140
2015	79 836	31 944
2016	81 024	32 412
2017	84 696	33 888
2018	89 940	35 976
2019	98 100	39 240

10.4. Categorisation of Self-Employment as a Primary or Secondary Gainful Activity in Relation to the Reason for Self-employment as a Secondary Gainful Activity

During the analysis, it was discovered that available data does not include information that could be used for clear categorisation of self-employment performed as either a primary or secondary gainful activity, see *"Conditions for Registration to Self-Employment as a Secondary Gainful Activity" above* – i.e., information such as for example whether a person receives parental allowance, is a student, or

took care of a dependent person. The most promising for this categorisation appeared an option sort data according to the amount of determined assessment base and the minimum assessment base limit for self-employment as a primary and secondary gainful activity. The determined assessment base amount, however, is not a reliable indicator because from the data it is not possible to find out whether the person performed only one type of self-employment during the entire year or only for a certain part of the year.

Even though, a rough division to primary and secondary activity, according to the determined assessment base and minimum assessment base, approximately corresponded with data from The Czech Social Security Administration (ČSSZ) (ČSÚ, 2021), it is not possible to find this division completely reliable, neither would be modelling further self-employment development simulations. Categorisation of probabilities of becoming self-employed and being self-employed as a primary or secondary gainful activity will be possible if necessary data is added (either in the form of direct identification of primary or secondary gainful activity or by adding information which could be used for identifying reasons for performing self-employment as a secondary gainful activity.

Also, during modelling the development of occupation / self-employment, self-employment as a primary or secondary gainful activity will be subsequently differentiated. The differentiation will occur because people will be assigned a status (student, care giver, etc.) during modelling - this value leads to differentiation between performing self-employment as a primary or secondary gainful activity.

10.5. Incorporation of the Topic of Self-Employment Before Starting the Project

So far, the model simulated situation whether a person will be employed. This covered both employees and self-employed without any clear differentiation. In case that a person became self-employed for some time, this was reflected only in changes in the amount of their assessment base resulting from an entered coefficient. Composing this chapter helped to specify the self-employment status more closely, to differentiate it unambiguously from an employee status and it also affected other composed topics.

10.6. Factors that Impact Self-Employment

Database VZ STATMIN OSVČ was used as the main source of information about self-employed persons. Database VZ STATMIN OSVČ includes data about self-employed persons who have paid pension insurance since year 2013 until year 2019. Impact of factors age and gender were analysed regarding their impact on the process of commencement or termination of self-employment.

10.6.1 Age

Data Availability

Age is one of the factors that were analysed in regard to the probability to commence or terminate self-employment. Information about age is available for each person's ID in all data sources.

Factor Impact

Basic analysis shows that the number of people who commence or terminate self-employment changes quite strongly with age. Age is a significant factor influencing the probability of becoming self-employed, becoming an employee, or performing concurrent activities. The most frequent self-

employment commencement occurs among people between 25 and 41 years of age. The number of people commencing self-employment decreases with increasing age.

10.6.2 Gender

Data Availability

Gender is another fundamental factor which was used during the analysis and probabilities calculations. Information about gender is available for each person's ID in all data sources.

Factor Impact

Differentiation between a self-employed man or woman influences both the absolute number of people and self-employment commencement / termination probabilities, in a similar manner as described for the age factor. Men comprise 70% of all self-employment commencement and termination records and women the remaining 30 percent.

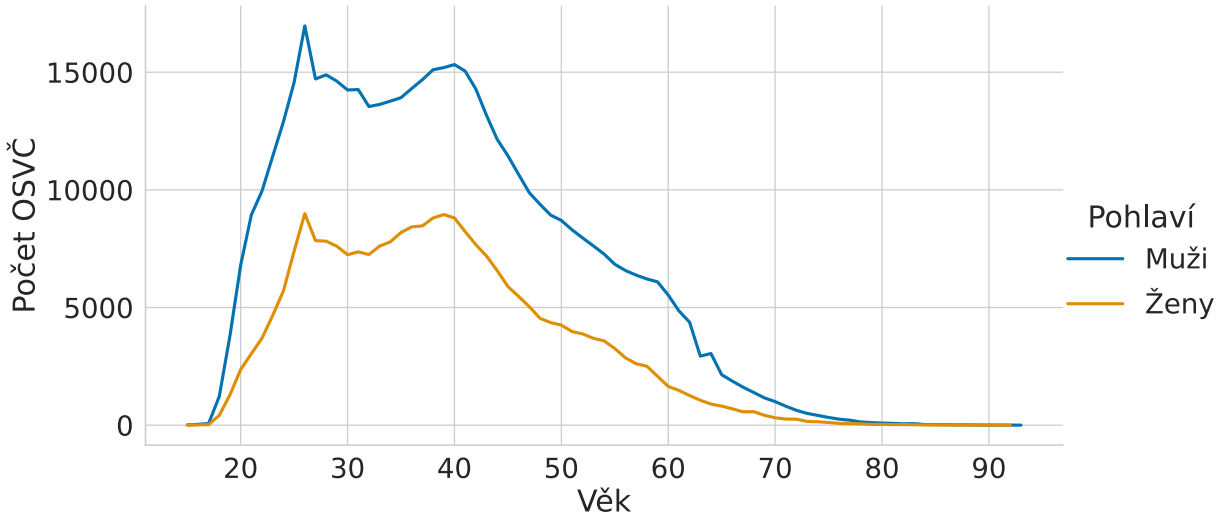

Figure 10.1: The number of people who commence self-employment dependent

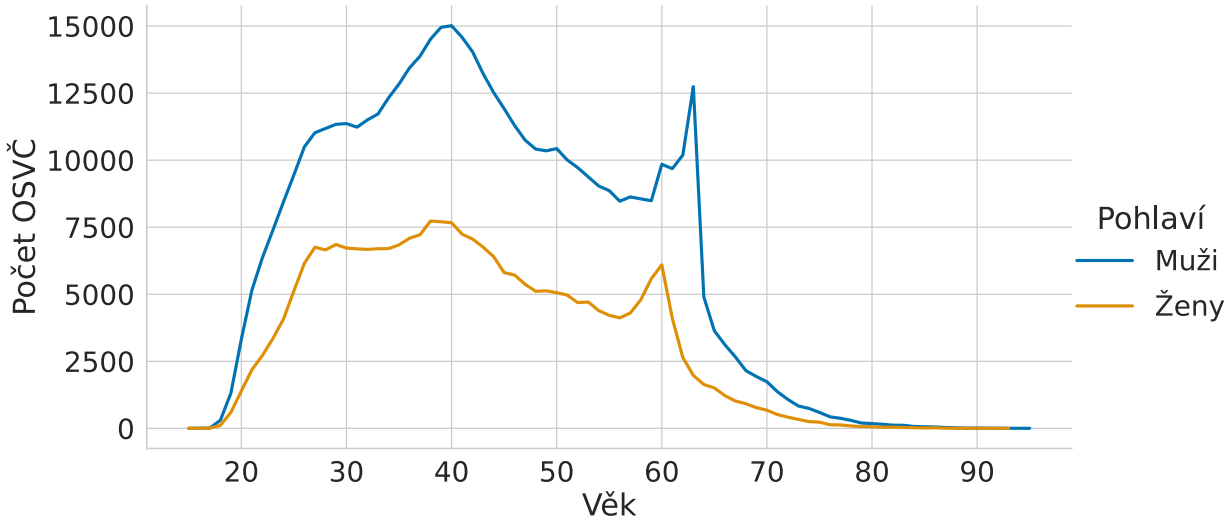

Figure 10.2: The number of people who terminate self-employment dependent

10.6.3 Other Factors

Also factors occupation and education were analysed – relevant information about education and occupation, however, is not available in database VZ STATMIN OSVČ. This information can be linked from the extended VZ STATMIN database in such a case when the given self-employed person was also employed at some point during their lifetime. That represents only 16% of unique IDs from the database VZ STATMIN OSVČ. It would not result in relevant information for occupation because the nature of some occupation can make them combinable with employment and self-employment or on the other hand also completely the opposite, i.e., concurrence might be impossible. Information about level of education is only available for 16% of all unique IDs present in database VZ STATMIN OSVČ. We would recommend including these factors in probability calculations if data about occupation, more precisely about professional fields of self-employed persons, and also data about education were directly available in database VZ STATMIN OSVČ.

10.7. Analysis of Possible Concurrence of Self-Employment with Employment and Probability Calculations

It was necessary to carry out multiple input data transformations for a more detailed analysis and calculations of probability that a person becomes an employee or self-employed. Input data consisted of databases VZ STATMIN and VZ STATMIN OSVČ. First step was to exclude deceased persons and records which show *null* assessment base for a given person and year (records which show *null* value of the assessment base in database VZ STATMIN OSVČ are such records where the given person owes money on advance insurance payments in the given year).

The last transformation is a decomposition of individual entries to monthly level because not every person was self-employed or employed for the duration of the whole year in that given year. Monthly granularity of records enables more detailed observations of transitions between employment and self-employment. There can be multiple entries for each ID in the given year (depending on the number of employments in database VZ STATMIN or on interrupting and resuming self-employment in the given year). In such a case, multiple lines were joined into one entry which, however, keeps information about the duration of individual employments and assessment bases.

Subsequently, both the VZ STATMIN database and database VZ STATMIN OSVČ are joined in order to enable observations whether in a given month, a given person is employed, self-employed or is in concurrence of self-employment and employment. Whether self-employment is commenced, terminated or whether there is a concurrence of both types of gainful activities is labelled on monthly bases. Probabilities are calculated separately thereafter for the following situations:

10.7.1 Change of Gainful Activity while on Labour Market

Changes in gainful activity are defined in the following combinations: situation when a person in status 11 (Employed) remains in this activity (labelled as OSVC_OSVC, EMPL_EMPL a BOTH_BOTH where first part of the label marks the original activity type and the second part after the underscore character marks the final activity type: EMPL stands for occupation, OSVC for self-employment and BOTH concurrence of self-employment and employment); or when a person in status 11 (Employed) transitions into the other gainful activity (i.e., OSVC_EMPL, EMPL_OSVC, BOTH_EMPL or BOTH_OSVC); or when they commence concurrence of these activities (i.e., EMPL_BOTH, OSVC_BOTH). The sum of probabilities for the same default position (e.g., OSVC_OSVC, OSVC_EMPL, OSVC_BOTH) gives 1 (i.e., 100%).

Graphs below show probabilities of transitions from one type of activity into a new one, separately for women and men. (Probability of remaining in the original activity is approximately 99% for each combination and it is not displayed in order to maintain clarity of the graph.)

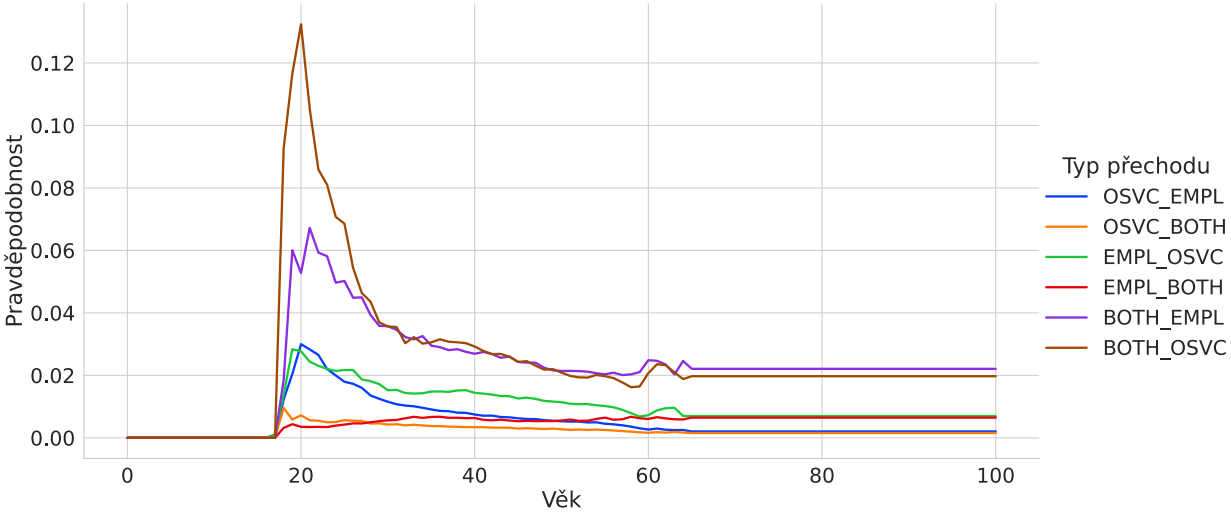

Figure 10.3: Probability of women making a change in their gainful activity displayed for different types of original activities, distributed by women's age.

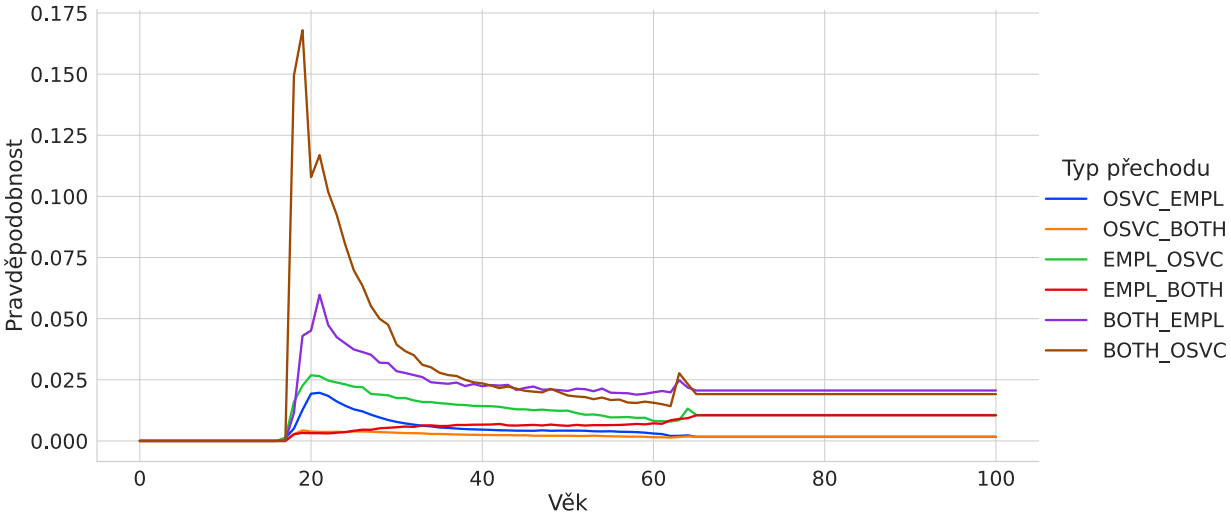

Figure 10.4: Probability of men making a change in their gainful activity displayed for different types of original activities, distributed by men's age.

Probabilities are calculated for persons at 15 years of age and above (as that is the legal age of entering the labour market, i.e., becoming an employee or a self-employed person) up to 100 years of age (age factor), calculated separately for men and women (factor gender). For persons aged 65 and above, a constant probability is used – calculated as a weighted average of individual probabilities from 65 years of age (including) until 100 years of age (including) separately for men and women. Weighted average was selected for the reason of low number of records for people above 65 years of age and using the standard method of calculation as for previous years resulted in significant fluctuations.

Probability of transition from one gainful activity to another is much more significant for men than for women. Figure 10.3 shows that most changes for women happen at the early stages of engaging in a gainful activity, specifically between 20 and 30 years of age when the probability of change is the highest for transitions to self-employment or to employment. Figure 10.4 shows that a similar trend holds for men. Probability of transition from one gainful activity to another for men who are still economically active (i.e., in status 11) does not change substantially change with increasing age, with the exception of time when old-age pension age is reached.

10.7.2 Change of Gainful Activity after Return to Labour Market from Unemployment or Inactivity

This a situation when a person interrupts their engagement in gainful activity (is assigned status 21 (Unemployed) or status 31 (Inactive)) and subsequently returns to status 11 (Employed). In such a case, probabilities of transition are calculated similarly as in the previous section. The sum of probabilities for the same default position (e.g., OSVC_OSVC, OSVC_EMPL, OSVC_BOTH) gives 1 (i.e., 100%). Factors age and gender are approached in a similar manner as in the previous sub-chapter "Change of Gainful Activity while on Labour Market".

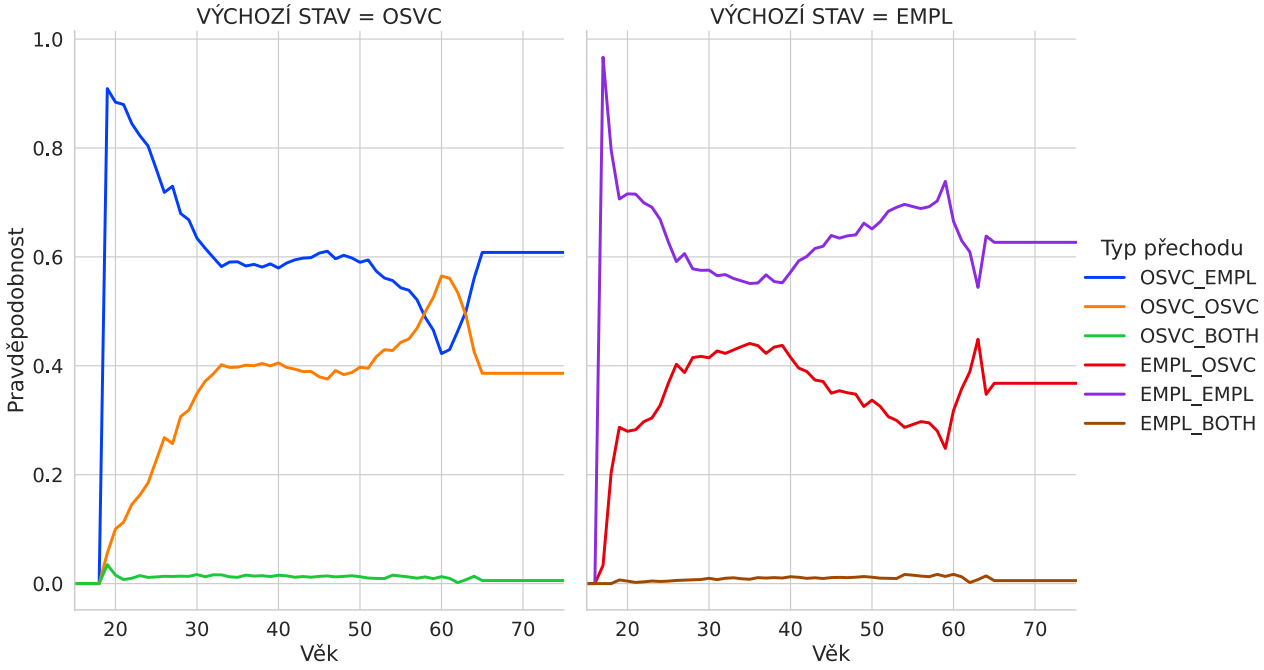

Figure 10.5: Probability of employment or self-employment commencement after return to the labour market, compared with a women's original choice, distributed by age

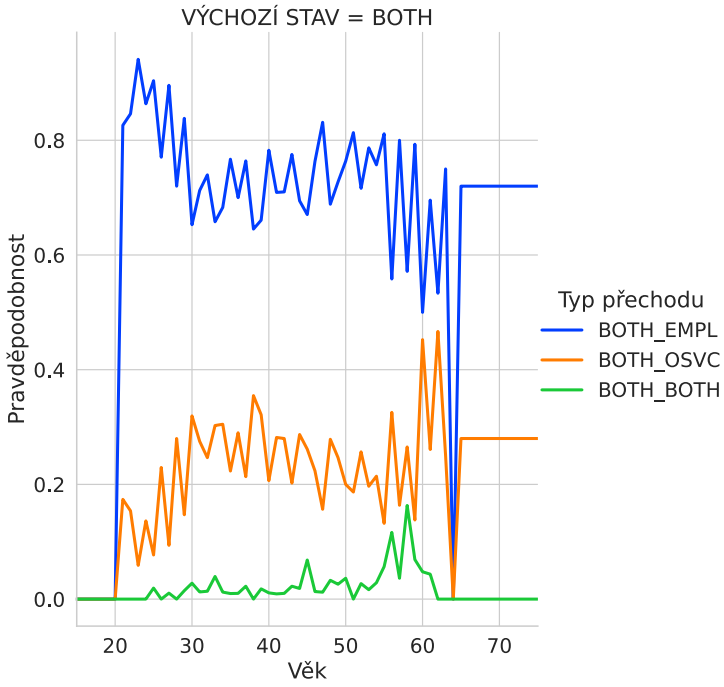

Figure 10.6: Probability of commencement of gainful activity concurrence after return to the labour market, compared with a women's original choice, distributed by age

In case of probabilities of return from inactivity to the labour market originally from self-employment, we can see a similar tendency of both men and women to re-enter self-employment with increasing age (Figure 10.5– initially self-employed = OSVC / Figure 10.7 – initially self-employed = OSVC, orange line).

Adequately, returns to labour market from employment (Figure 10.5 and Figure 10.7, initially employed = EMPL) show as most significant to find employment again (i.e., not to enter concurrence or become self-employed). An exception is the old-age pension commencement period when a growing probability for both men and women is that they will become self-employed – approximately from 60 years of age.

In case of returning to the labour market from previous concurrence of gainful activities Figure 10.6 and Figure 10.8, initially in concurrence = BOTH), the highest probability is that the person will become an employee (probability fluctuates between approximately 60 and 80% due to low number of events, dependent on age), followed by probability that the person will become self-employed (35% to 45%). The lowest probability here is that they will re-enter concurrence of gainful activities.

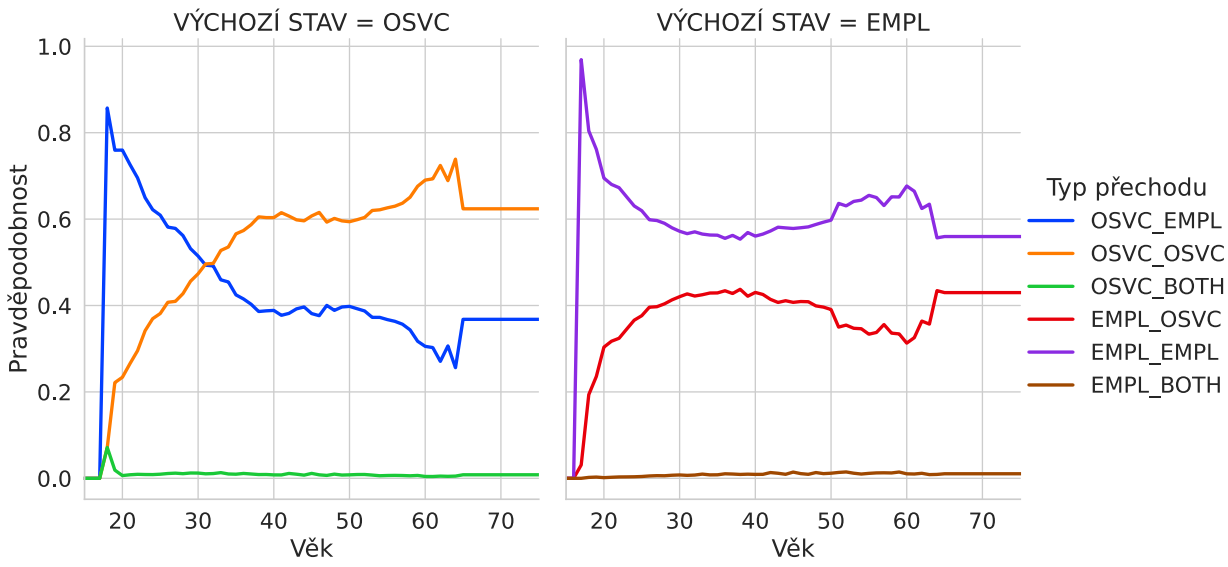

Figure 10.7: Probability of employment, self-employment and commencement of gainful activity concurrence after return to the labour market, compared with a men's original choice, distributed by age

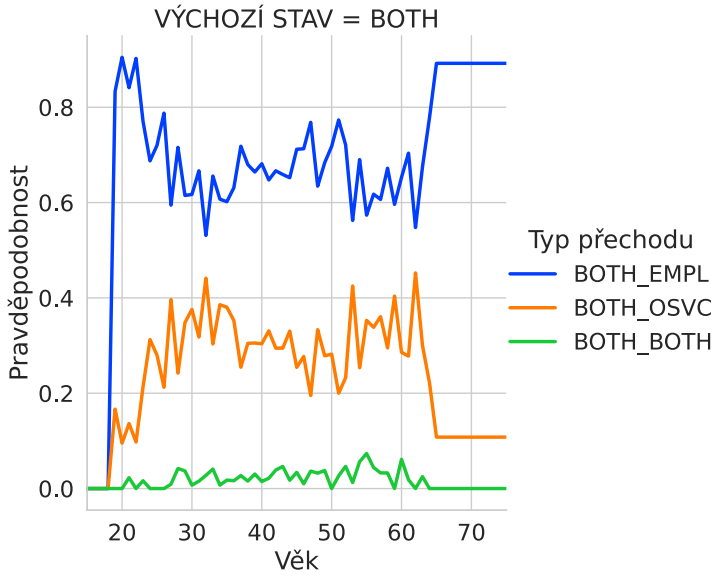

Figure 10.8: Probability of commencement of gainful activity concurrence after return to the labour market, compared with a men's original choice, distributed by age

10.8. Analysis of Return to Self-Employment after Having Been Self-Employed in Previous Years

A sub-analysis of self-employment describes the probability of return to self-employment in case that a given person has been self-employed previously. The probability of first-time commencement of self-employment has been added for illustration. Data from database VZ STATMIN OSVČ and database INEP were used for this analysis.

The line chart (Figure 10.9) shows monthly probability of return to self-employment (a given person has been self-employed before) on dashed lines – it is apparent that the probability of return to self-employment increases with increasing age up to approximately 40 years of age and then it starts to drop until retirement age. The graph also illustrates monthly probability of first-time self-employment commencement (solid lines) – probability that a person commences self-employment for the first time significantly decreases with increasing age.

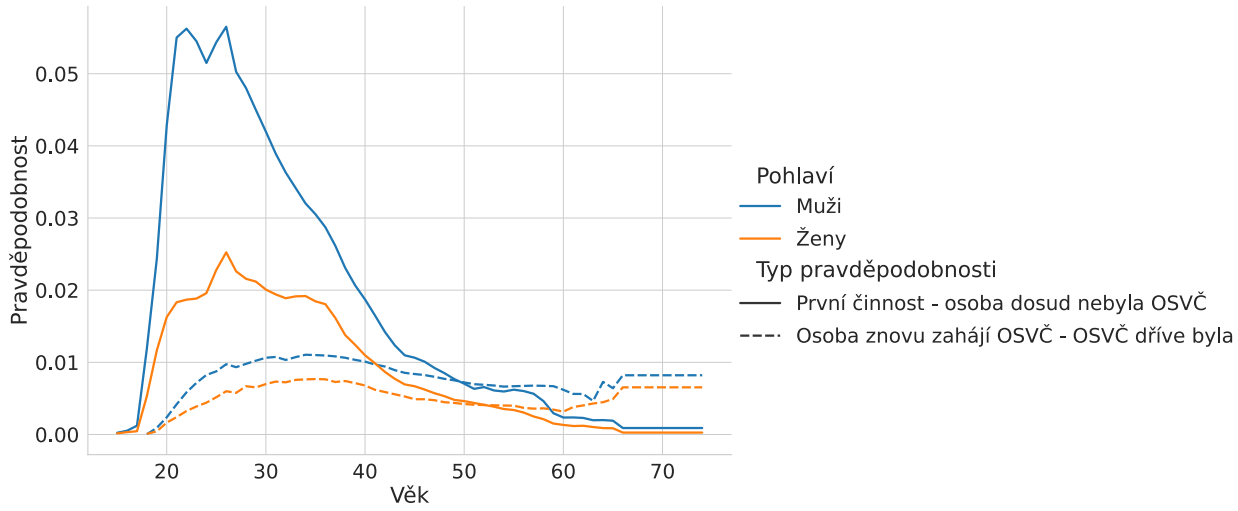

Figure 10.9: Probability of self-employment commencement in case that a given person has already been previously self-employed and an additional probability of commencing self-employment for the first time

Following analysis (Figure 10.10) compares the number of people commencing self-employment for the first time with the number of people of all persons commencing self-employment. The analysed period was from year 2013 until year 2019. The analysis shows that, in the specified timeframe from year 2013 until 2019, a significant proportion of persons commencing self-employment after reaching 40 years of age have already been self-employed previously (blue dashed line in the line chart below) and this trend is even more signified for self-employed persons after their 50 years of age.

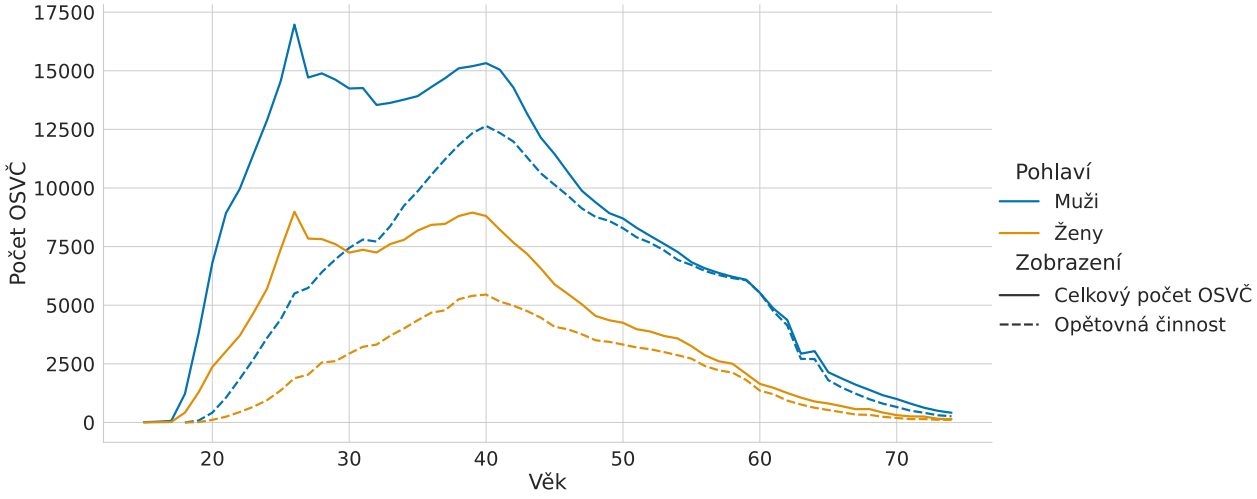

Figure 10.10: The number of persons commencing self-employment – differentiation between the total number of self-employed persons and those returning to self-employment who have previously been self-employed

10.9. Implemented Changes

10.9.1 Preparation of MPs

Analogically to the previous chapters, it is necessary to determine for each individual in the model point database whether it will be entering the model as an employee or as self-employed (`INIT_EMPL, INIT_OSVC`) and also their history of assessment bases `(HIST_GRS_SAL_EMPL` or `HIST_GRS_BASE_OSVC`). Data used for this purpose were taken from updated databases INEP_PARTICIPANTS, VZ, VZ OSVC and the last entry about the given person from year 2019 helped determine whether this person will become an employee or self-employed (`INIT_EMPL` or `INIT_OSVC`). The model point table was extended to include self-employment and columns `HIST_GRS_SAL_EMPL and HIST_GRS_BASE_OSVC` with suffixes 1 to 34 where 1 marks the column with assessment base from year 2019 and column 34 carries value from year 1986 (for example `HIST_GRS_SAL_EMPL_1`).

Input columns newly added to Prophet are here column `INIT_EMPLOYEE`, column field `HIST_GRS_SAL_EMPL` and column field `HIST_GRS_BASE_OSVC`. These columns first get added to the end of database `INEP_PARTICIPANTS`. This happens when starting DCS `MERGE_INEP_EXT_INPUTS`. Input file including personal IDs, corresponding indicator `INIT_EMPL` / `INIT_OSVC` and column with income / assessment base history `HIST_GRS_SAL_EMPL` / `HIST_GRS_BASE_OSVC`, must be placed into directory `INPUTS/INEP_EXT_INP_OSVC.csv`.

While extending INEP_PARTICIPANTS to include the above-mentioned columns, a check is carried out for each person to determine whether each person has been assigned at least one from tags `INIT_OSVC` and `INIT_EMPLOYEE` (it should never be the case that a person would be neither employed nor self-employed). In case that such a person exists in the dataset, what happens first is comparison of each assessment base year – if earlier entry exists in field `HIST_GRS_SAL_EMPL`, the person gets assigned tag `INIT_EMPLOYEE = 1`. In case that such a person has an earlier entry in field `HIST_GRS_BASE_OSVC`, they will be marked as self-employed (`INIT_OSVC = 1`). If the person does not have any assessment base history as either self-employed nor as an employee, the decision whether they will be marked as employee or self-employed is dictated by the probability given by FAC table `empl_osvc_prob.fac`.

In most programmes in DCS, the only change was in input and output formats. Specifically, in case of adding input columns `INIT_EMPLOYEE`, `HIST_GRS_SAL_EMPL` and `HIST_GRS_BASE_OSVC` to input format `INEP_modelpoints`, and their addition to the output format `Modelpoints`. For correct transformation of input columns to output columns, reconstruction of both fields `HIST_GRS_SAL_EMPL[n]` and `HIST_GRS_BASE_OSVC[n]` must occur in code of all DCS programmes, that is from input columns `HIST_GRS_SAL_EMPL _[1 … n]` and `HIST_GRS_BASE_OSVC_[1 … n]`.

Further smaller changes in code were carried out for newly generated persons (children and immigrants) (`03_newborn_and_children.DCS` and `04_immigrants.DCS`). In both cases the assignation of either self-employed or employee is decided based on probability. FAC tables `births_osvc_empl.fac` and `immigrants_osvc_empl.fac` are used to determine these indicators. Initial assessment base (`INIT_GRS_SAL_EMPL` or `INIT_GRS_BASE_OSVC`) is then entered. Fields `HIST_GRS_SAL_EMPL` and `HIST_GRS_BASE_OSVC` remain at zero value for these types of newly generated persons.

10.9.2 Prophet

Changes in the model can be divided into four categories:

New status EMPLOYEE

New status `EMPLOYEE` under status 11 (employed) was added into table `status_vars.fac`. Status `EMPLOYEE` enables modelling transitions between an employee and a self-employed person, see next bullet point.

Transition between status EMPLOYEE and status OSVC (self-employed) (and their combinations)

Event `Employed_change` is used for modelling this transition. Probabilities of the transition depend on status (`EMPLOYEE, OSVC`) in the last month, also on gender and age. These probabilities can be found in table `Employed_change.fac` which is described below. Changes are modelled in status employed (11) and revised at transition from status unemployed / inactive (21/31) to status employed (11).

Addition of a new stochastic event means increase in the value of variable `NO_EVENTS` in table `Global.fac` by 1. Extraordinarily, it was necessary to remove some existing events in this chapter, specifically events `OSVC_Start` and `OSVC_Stop`, due to change in self-employment concept in the model. Altogether, `NO_EVENTS` was decreased by 1. Removing events `OSVC_Start` and `OSVC_Stop` resulted in alterations of variable which were used in these events and relevant probabilities were deleted from tables.

Self-Employment as Secondary Gainful Activity

An option of self-employment as secondary gainful activity was added into the model. This type of self-employment is bound by fulfilment of requirements, this is specified by variable `SUB_OSVC_ELIG`. Requirements for self-employment as secondary gainful activity are defined in table `sub_OSVC.fac` (see table descriptions below).

Variables for Self-Employment

Variables regarding the following topics were set for self-employment:

- assessment base, assessment base history
- gross income, net income
- advance payments towards pension, health and sickness insurance
- sick pay, maternity leave payments and unemployment benefits
- pillar 3 (and potentially pillar 2) set the same ways as for employees, with the exception of employer's contribution.

These variables were for status `EMPLOYEE` only revised.

In that regard, table `contribution_rates.fac` was extended to include percentage of contribution to governmental employment policy.

10.9.3 Description of .fac tables

Employed_change.fac

This table determines monthly probability of transition between status EMPLOYED and status OSVC (self-employed) (and their combinations). It was created by merging and modifying the structure of the tables `Employed_change_EA_11.csv` containing the probabilities of change without interruption of economic activity and `Employed_change_EA_21_31.csv` containing the probabilities of change with interruption of economic activity.

Table 8.8: Structure of Table Employed_change.fac

Code	Comment
SEX	Gender
AGE_NOW	Current age
EMPLOYEE_OLD	Initial status EMPLOYEE (1/0)
OSVC_OLD	Initial status OSVC (1/0)
EMPLOYEE_NEW	New status EMPLOYEE (1/0)
OSVC_NEW	New status OSVC (1/0)
CHANGE_PROB	Probability of transition
CHANGE_PROB_NO_EVENT	Probability of transition without event

sub_OSVC.fac

This table determines the entitlement to self-employment as a secondary gainful activity according to the status. Each status can be found in columns (`STUDENT, DISABLED` and the like). Their values determine whether the given status gives entitlement to self-employment as a secondary gainful activity (1 = entitled).

Macro_65_short_proj_updateCSU2018_10_6.fac

This table is used for generating macroeconomic scenarios. Only added columns are described. Numbers in the columns stated below that end `_BASE` denote percentage of national average salary for a given year.

Table 8.9: Structure of Table Macro_65_short_proj_updateCSU2018_10_6.fac

Code	Comment
SUB_OSVC_MIN_BASE	Minimum assessment base for self-employment as a secondary gainful activity, for retirement
OSVC_MIN_HEALTH_BASE	Minimum assessment base for self-employment as a primary gainful activity, for health insurance
SUB_OSVC_MIN_HEALTH_BASE	Minimum assessment base for self-employment as a secondary gainful activity, for health insurance
MIN_DECISIVE_INCOME	Minimum decisive income (for sickness insurance)

Empl_osvc_prob.fac, Births_osvc_empl.fac a Immigrants_osvc_empl.fac

Table `empl_osvc_prob.fac` is used to determine whether an existing person is an employee or self-employed in such cases when there are no records in assessment base history in `INEP_PARTICIPANS`.

Tables `births_osvc_empl.fac` and `immigrants_osvc_empl.fac` are used to determine whether a newborn, child or immigrant is an employee or self-employed.

Table 8.10: Empl_osvc_prob.fac, Births_osvc_empl.fac and Immigrants_osvc_empl.fac Table Structure

Code	Comment
SEX	Gender (table immigrants_osvc_empl.fac does not distinguish between genders)
PROB_EMPL	Probability that a person of the given gender will become an employee
PROB_OSVC	Probability that a person of the given gender will become self-employed

10.10. Implementation Feasibility Assessment of Other Factors

Difficulty of implementation of additional factors (education, occupation should be low. It would consist primarily of edits in table *`employed_change.fac`* where it would be necessary to add another column (columns) including data about the additional factors and determining probabilities of transition between *`OSVC, EMPLOYEE`* and their combinations.

We see a potential problem in regard to longer time needed for calculations – tables are already quite large. We see a potential problem in regard to longer time needed for calculations – tables are already quite large. Adding additional factors (if they proved significant in the future or if better quality data was available) would result in several times larger table sizes. Therefore, it is important to be wary of the effect on calculation times in case of a potential implementation of additional factors.

10.11. Suitability of Future Implementation of Other Factors

Table 8.11: Suitability of future implementation of additional factors

Factor	Data Availability	Significance	Implementation Difficulty	Recommendation
Age	Excellent	High – age impacts probability whether a person commences or terminates self-employment or employment	Implemented	Implemented
Gender	Excellent	High – men become self-employed more often than women	Implemented	Implemented
Education	Poor, in the extended VZ STATMIN database there is available data about 16% IDs from VZ OSVČ	Low – analysis was not carried out due to lack of data	Cannot be determined	We recommend for consideration in case that data becomes available
Occupation	Poor, in the extended VZ STATMIN database there is available data about 16% IDs from VZ OSVČ	Low – analysis was not carried out due to lack of data	Cannot be determined	We recommend for consideration in case that data becomes available

11. Attachments

A. List of Abbreviations

CBOLT	Congressional Budget Office Long-Term
ČSSZ	Česká správa sociálního zabezpečení (Czech Social Security Administration)
ČSÚ	Český statistický úřad (Czech Statistical Office)
CSV	File format .csv (Comma Separated Values)
CZ-ISCO	Klasifikace zaměstnání ISCO (International Standard Classification of Occupations, Czech revision)
DCS	Prophet Data Conversion System
DELNEZ	A tool for probabilistic pairing between the STATMIN VZ and SEE databases.
DPČ	Dohoda o pracovní činnosti (Agreement to perform work)
ELDP	Evidenční list důchodového pojištění (Personal records for pension insurance)
ID	Unique personal identifier within a database
INEP_PARTICIPANTS	Database of modelpoints entering the DCS tool
ISPV	Informační systém o průměrném výdělku (Average income information system, system of income monitoring according to the EU Structure of Earnings Survey)
KVČ	Kód výdělečné činnosti (gainful activity code)
LAU	Local administrative unit (místní správní jednotka)
MP	Model point
MPSV	Ministerstvo práce a sociálních věcí (Ministry of Labour and Social Affairs)
NEMO	Dynamic microsimulation pension model of the Ministry of Labour and Social Affairs

NUTS	Nomenclature des Unités Territoriales Statistiques (Nomenclature of Territorial Units for Statistics)
OSVČ	Osoba samostatně výdělečně činná (a self-employed person)
PED	Permanent Earnings Differential
PnP	Příspěvek na Péči (care allowance)
PSČ	Poštovní směrovací číslo (postal code)
RMSE	Root Mean Square Error
VZ	Vyměřovací základ (assessment base)

B. Bibliography

Congressional Budget Office. (2013). Modeling Individual Earnings in CBO's Long-Term Microsimulation Model. Washington, D.C.

Česká pošta. (2021). *Seznam části obcí a obcí s adresním PSČ*. Retrieved from https://www.ceskaposta.cz/documents/10180/3738087/xls_cobce_psc.zip/9cdf9b25-23c4-206e-fb58-b28430b49ff6

Český statistický úřad. (2021). *Číselník okresů*. Retrieved from https://data.cssz.cz/tabulka-ciselnik-okresu

ČR, P. s. (2021, 09 20). *Předpis 50/1978 Sb. Vyhlášeka Českého úřadu bezpečnosti práce a Českého baňské úřadu o odborné způsobilosti v elektrotechnice*. Retrieved from https://www.psp.cz/sqw/sbirka.sqw?cz=50&r=1978

ČSÚ. (2018, 02 09). *Struktura věty RES pro externí uživatele.* Retrieved from https://www.czso.cz/csu/res/struktura_vety_res_pro_externi_uzivatele

ČSÚ. (2020, 06 23). *Klasifikace zaměstnání (CZ-ISCO)*. Retrieved from https://www.czso.cz/csu/czso/klasifikace_zamestnani_-cz_isco-

ČSÚ. (2021). *Číselník okresů*. Retrieved from http://apl.czso.cz/iSMS/cisdata.jsp?kodcis=109

ČSÚ. (2021, 07 01). *Definice a druhy SVČ*. Retrieved from https://www.cssz.cz/definice-a-druhy-svc

ČSÚ. (2021, 07 15). *Graf počtu OSVČ v ČR*. Retrieved from https://data.cssz.cz/web/otevrena-data/graf-pocet-osvc-v-cr

ČSÚ. (2021). *Počet obyvatel v regionech soudržnosti, krajích a okresech České republiky k 1. 1. 2021*. Retrieved from https://www.czso.cz/csu/czso/pocet-obyvatel-v-obcich-k-112021

ČSÚ. (2021, 02 01). *Seznam částí obcí a obcí s adresním PSČ, Česká pošta*. Retrieved from https://www.ceskaposta.cz/ke-stazeni/zakaznicke-vystupy

ČSÚ. (2021). *Územní, sídelní struktura*. Retrieved from https://vdb.czso.cz/vdbvo2/faces/shortUrl?su=ae3f1b9a

DataSentics. (2019). *Tvorba podkladové databáze identifikátorů osob k analýze dat o dočasné pracovní neschopnosti a invaliditě.*

Deloitte. (2014). *Studie proveditelnosti – Implementace rozhodovacích procesů do dynamického mikrosimulačního modelu důchodového systém.* Praha: .

Národní ústav pro vzdělávání. (2021, 09 20). Retrieved from http://www.nuv.cz/t/rv

Plívová, V. (2011). Nový statistický model důchodu umí využít invidividuální data. Praha. Retrieved from https://www.mpsv.cz/documents/20142/805559/06052011.pdf/d32c083d-73ae-ce0c-1442-661eabcc606b

Šlapák, Holub, Průša. (2017). Metodika: Identifikace sociodemografických charakteristik ovlivňujících načasování odchodu do důchodu.

Trexima. (2015). Odvození parametrů rovnice mzdové dynamiky pro dynamický mikrosimulační model důchodového systému MPSV. Zlín - Louky.

TREXIMA. (2019). *Informační systém o průměrném výdělku.* Retrieved from https://www.ispv.cz/cz/Vysledky-setreni/Archiv/2019.aspx

C. List of Tables

D. List of Figures