sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 3
ISSN 1507-3866; e-ISSN 2449-9994

GRAPHICAL TOOLS OF DISCRETE LONGITUDINAL DATA PRESENTATION IN R*

Ewa Genge
University of Economics in Katowice, Katowice, Poland
e-mail: ewa.genge@ue.katowice.pl

ORCID: 0000-0002-8899-3697

© 2019 Ewa Genge
This is an open access article distributed under the Creative Commons Attribution-NonCommercial-NoDerivs license (http://creativecommons.org/licenses/by-nc-nd/3.0/)

DOI: 10.15611/eada.2019.3.03
JEL Classification: C00, G40

Abstract: Good graphical presentation of data is useful during the whole analysis process from the first glimpse into the data to the model fitting and presentation of results. The most popular way of longitudinal data presentation are separate (for each wave, in cross-sectional dimension) comparisons of figures. However, plotting the data over time is useful in suggesting appropriate modeling techniques to deal with the heterogeneity observed in the trajectories. The main aim of this paper is to present the changing perceptions of the financial situation in Poland using different graphical tools for the heterogonous discrete longitudinal data sets and present demographics features for those changes. We will focus on the most important features of the categorical longitudinal data – category sequences and their graphical presentation. We aim to characterize the analyzed sequences on the basis of unidimensional indicators and composite complexity measures, as well as using mainly `TraMineR` [Gabadinho et al. 2017] package of R.

Keywords: longitudinal data, categorical sequences, sequence visualization.

1. Introduction

One of the first, but still convincing methods of describing the structures in data is visualization. Graphical data presentation is not only data depiction but it is frequently considered as the basic method of data analysis. This approach is very popular in analyses of bi-dimensional data sets, which can be presented on the surface and submitted to the initial visual inspection of the whole data set.

Good graphical presentations of data are useful during the whole analysis process from the first glimpse into the data to the model fitting and presentation of results.

* The author would like to acknowledge the research grant (SONATA 12, UMO-2016/23/D/HS4/00989, "Latent variable models in the identification of homogenous structures in socio-economic longitudinal data") of the National Science Centre, Poland.

This work is focused on discrete longitudinal (panel) dataset[1]. Longitudinal research can take numerous forms: 1) *repeated cross-sectional studies* where participants are largely or entirely different on each sampling occasion; 2) *prospective studies* where the same participants are followed over a period of time; 3) *retrospective studies* are designed after the experienced events.

The main types of prospective studies include *cohort panels* (where some or all individuals in a defined population share the same event are considered over time, i.e. a clinical cohort) and *representative panels* (where data is regularly collected for a random sample of a population), subject of this work.

The most popular way of longitudinal data presentation are separate (for each wave, in cross-sectional dimension) comparisons of figures. However, plotting the data over time is useful in suggesting appropriate modeling techniques, such as the growth mixture model [Muthén, Shedden 1999], to deal with the heterogeneity observed in the trajectories.

Longitudinal data are often visualized using a growth plot, also known as a growth curve or trajectory plot[2] [Singer, Willett 2003]. However, plotted growth curves for multiple participants rapidly become uninterpretable with categorical data (see e.g. [Tueller et al. 2016, Figure 1, p. 1]). Categorical data define specific levels (an arbitrary number of categories, e.g. employees, entrepreneurs, farmers, pupils and students, the unemployed), and these levels do not necessarily need to represent any hierarchical order. Thus, a trajectory becomes a sequence of categorical data (response configurations) rather than a continuum [Tueller et al. 2016].

The main aim of this paper is to present the changing perceptions of the financial situation in Poland using different graphical tools for the heterogonous discrete longitudinal data sets, and present the demographics features for those changes. The author will focus on the most important features of the categorical longitudinal data – category sequences and their graphical presentation (sequence index plot, sequence frequency plots, mean time plot, transversal distribution plot, and the transversal entropy plot). Furthermore, it is aimed to characterize the analyzed sequences on the basis of unidimensional indicators and also composite complexity measures.

2. Data

An analysis of income, the primary measurement of household's wealth, conducted by the Social Monitoring in the framework of the Social Diagnosis 2015 [Social Diagnosis 2015], shows that the financial situation of households in Poland has improved in recent years. It is interesting to identify the demographic characteristics

[1] A longitudinal (panel) dataset tracks the same type of information on the same subjects at multiple points in time.

[2] A scatter plot with time on the horizontal axis and the values of the response variable being studied on the vertical axis.

associated with the changes in the financial situation of Polish families. We will concentrate especially on the graphical presentation of those changes.

In the next sections we analyze the questionnaire point: "Is it easier to make ends meet?" measured by ordinal response variable (with great difficulty; with difficulty; with certain difficulty; rather easily; easily) for each wave of Social Diagnosis panel research (2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015). We also consider the covariates: socio-economic group (1 – employees in public and private sector, 2 – farmers, 3 – self-employed, 4 – retirees, pensioners, and living on unearned sources), family type (1 – married without children, 2 – married with one child, 3 – married with two children, 4 – married with three or more children, 5 – other family types (one-parent family, multi-family, non-family-one person, non-family-multi-person), size of the place of the residence (1 – cities with more than 500,000 inhabitants, 2 – cities with 20,000 to 500,00 inhabitants; 3 – cities below 20,000 inhabitants and rural areas). We also consider the survey weights that account for the sampling scheme and unit non-responses.

There are 346 complete observations at each point of time. In total there is information on $n = 2768$ cases. The public data set is available at http://www.diagnoza.com/index-en.html.

All computations and graphics in this paper have been done in `TraMineR` [Gabadinho et al. 2017] package of R.

3. Visualizing the individual sequence of categories

The sequence of categories (the response configuration) can be represented in many different ways, depending on the data source and on how the information is organized (long or wide format of the longitudinal data)[3]. In this section of the article we present and compare the results for the sequence index plot, sequence frequency plot, mean time plot, transversal distribution plot, and the transversal entropy plot for the subjective assessment of the financial situation of Polish households.

Sequence index plot presents the requested individual sequences which are rendered with horizontal stacked bars depicting the different categories over successive points of time.

Using `seqiplot()`[4] function of `TraMineR` [Gabadinho et al. 2017] package of R the individual longitudinal patterns as well as the duration spent in each of successive positions can be shown.

---

[3] Data organization and conversion between formats of longitudinal data is discussed in detail in [Ritschard et al. 2009].

[4] `seqiplot()`, as with most other plotting functions described in this paper, is just an alias for calling a generic `seqplot()` category sequence plot function with the suitable type argument and default option values. The structure of the general but complex plot function to be applied in R (preceded by the appropriate data preparation) is given by `seqdplot(seqdata, group = NULL, type = "", main = NULL, ...)`.

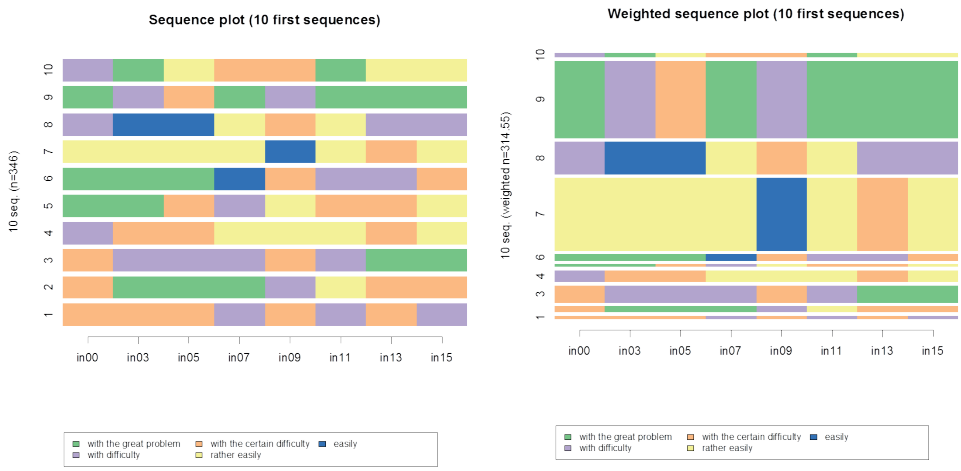

Fig. 1. Sequence index plot (unweighted and weighted) of sequences 1:10

Source: own calculations in R.

We presented the sequence index plot for the weighted[5] and unweighted frequencies and the first ten sequence frequencies (Figure 1), as well as the full index plots (Figure 2). In sequence number 1 (first one from the bottom in Figure 1) the respondent stayed in the same positions during the first three waves then three events occurred: he/she changed status between 2005 and 2007, 2009 and 2011, and again between 2013 and 2015 from "with certain difficulty" to "difficulty" status. The width of the bar representing each sequence (the right panel of Figure 1) is proportional to its weight.

The full index plots[6] display all the sequences in the set without spaces between sequences and without borders around the response categories. The benefit of such plots, for instance, was stressed by Scherer [2001] and Brzinsky-Fay et al. [2006]. However, when the number of displayed sequences is large, they may produce pictures that are often hard to interpret. A good choice, for instance, is the focus on the distance to the most frequent sequence (D/2-CD/1-RE/1-CD/3-RE/1, see Table 1) given at the bottom of the middle panel of Figure 2. Another choice is presenting the sequences sorted by the scores of a multidimensional scaling analysis (MDS) based on the dissimilarities between sequences, see also [Gabadinho et al. 2011, pp. 12-14] for more details. Then, at the bottom of the right panel of Figure 2 we can see the sequences with the lowest scores representing the worse financial situation (mainly "with great difficulty", "with difficulty" categories) as opposed to the sequences at the bottom of this figure.

[5] The y-axis indicates the cumulative percentage of the represented sequences.

[6] `seqIplot()` alias can be used to produce the full index plot.

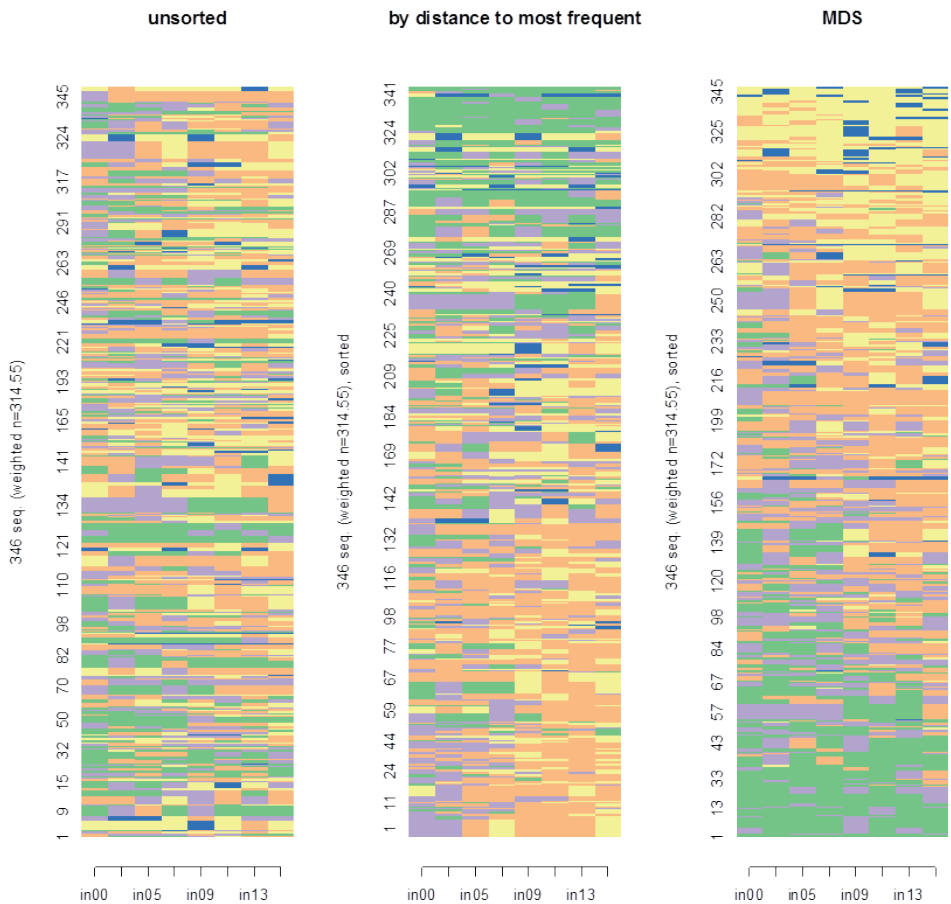

Fig. 2. Unsorted and sorted full-sequence index plots by the distance to the most frequent sequence and MDS

Source: own calculations in R.

Sequence frequency plot is obtained with `seqfplot()` function and is a graphical view of the sequence frequency tables (see Table 1)[7] where the bar widths are proportional to the frequencies. Figure 3 presents the plot of the weighted frequencies. The y-axis indicates the cumulative percentage of the represented sequences.

The most frequent sequence is to cope with the difficult financial situation during two waves followed by one wave of certain difficulty and rather easily and then once again 'certain difficulty' (for three waves) and then 'rather easily' level of income. The next three sequences in the weighted frequency table with 2.0%, 1.9% and 1.6% of the total weight reveal the improving financial situation in recent years.

[7] In Table 1 only the five most frequent sequences (in decreasing order) are shown.

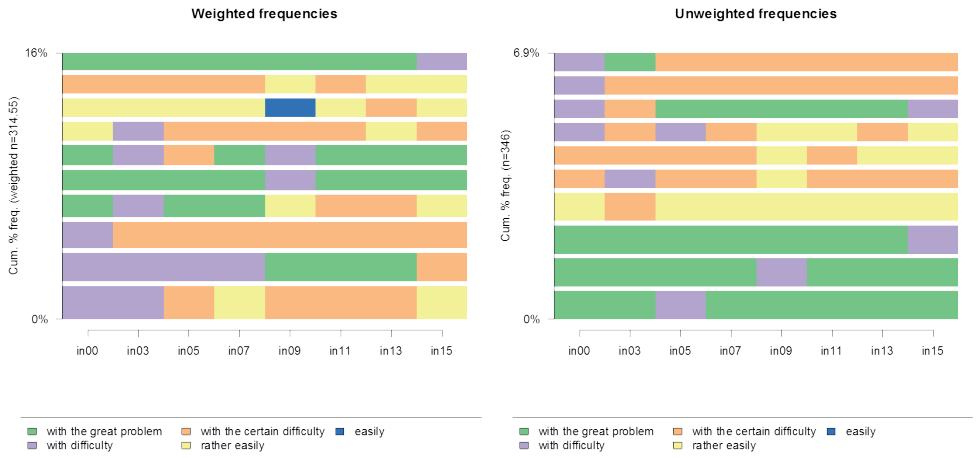

Fig. 3. Sequence frequency plot

Source: own calculations in R.

Table 1. Sequence frequency table (weighted frequencies)

State	Frequency	Percent
D/2-CD/1-RE/1-CD/3-RE/1	7.4	2.4
D/4-GP/3-CD/1	6.4	2.0
D/1-CD/7	5.9	1.9
GP/1-D/1-GP/2-RE/1-CD/2-RE/1	5.1	1.6
GP/4-D/1-GP/3	4.6	1.5

Source: own calculations in R.

Table 2. Sequence frequency table (unweighted frequencies)

State	Frequency	Percent
GP/7-D/1	3	0.87
RE/1-CD/1-RE/6	3	0.87
GP/2-D/1-GP/5	3	0.87
GP/4-D/1-GP/3	3	0.87
CD/1-D/1-CD/2-RE/1-CD/3	2	0.58

Source: own calculations in R.

The ten most frequent sequences account for only about 16% of all the trajectories, which reflects this high diversity.

As far as the unweighted frequencies are concerned, the studied households most frequently declared ‘great difficulty with making ends meet’ during seven waves followed by one wave of difficulty level of income, as well as the spell of one wave

with the ‘rather easily’ income’s level followed by one wave of ‘certain difficulty’ and then six waves of ‘rather easily’ levels of income assessment.

These two sequences, as well as the other two given below in the Table 2 which are the first four most frequent in the unweighted frequency table, each with 0.87% of the 346 cases considered, yielding a large number of different patterns.

It should be noticed that the sequence index plots and sequence frequency plots, do render individual sequences or individual follow-ups and can also be presented using functions of `longCatEDA` [Tueller 2017] and `SeqHMM` [Helske, Helske 2017] packages of R. In the next section the overall and transversal descriptive statistics of a set of sequences are presented.

4. Plotting overall and transversal statistics

Mean time plot displays the mean number of times each response category is observed in the sequence. We illustrated the results by socio-economic features such as family type, socio-economic group and size of the place of residence.

We can see (Figure 4) that the mean time for the households ‘coped with great difficulty’ is the highest for the other family types (2.14) and those with more children (1.65), farmers (2.7), living in small cities and rural areas (1.18) while the mean time for declaring easily making ends meet is higher for households of employees (0.34) with no children (0.93), living in big cities (0.38).[8]

Transversal distribution plot can be generated using `seqdplot()` and represents the sequence of the cross-sectional category frequencies by position (points of time). This plot provides the aggregated view and displays the general pattern of the whole set of trajectories.

The results shown in Figure 5 (presenting the distribution for different covariates) exhibit significant jumps especially for ‘no children’, ‘self-employed’ and ‘living in big cities’ family types. The highest and the increased proportion of the best income’s level assessment (‘easily’) at the end of the period could be observed for ‘no children’ families, employees, self-employed and people living in big cities. However, this kind of financial situation assessment does not occur among farmers at all. We can also observe the decreasing proportion of the lowest income declaration for all of the family types, size of place of residence as well as socio-economic groups at the end of the analyzed period of time. More details can be given on the basis of the transversal entropy plot (see Figure 6).

[8] ***Modal sequence plot.*** An interesting summary is also a presentation of the most frequent category at each point of time (modal sequence plot). Due to limitations of space we do not present the figures, we only mention that e.g. the most frequent sequence for the ‘no children family’ type is coping with ‘certain difficulty’ during three waves followed by one wave of difficulty declaration and again three waves of difficulty declaration, then followed by one wave of ‘rather easily’ declaration.

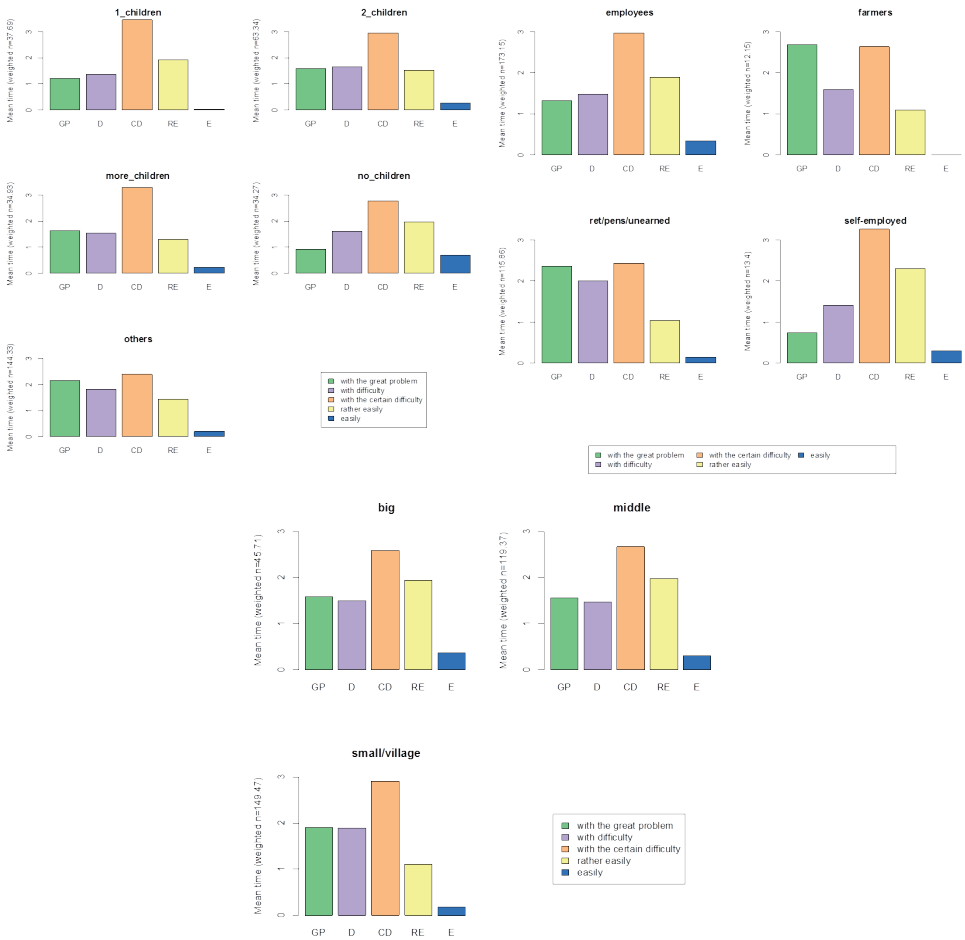

Fig. 4. Mean time plot by type of family, socio-economic group of household and place of living covariates

Source: own calculations in R.

Transversal entropy plot can be presented using `seqHplot()` function of R and displays the evolution over positions of the transversal entropies [Billari 2001]. The graphical presentation of the transversal entropy (known also as entropy index) can be useful to find out how the diversity of categories evolves along the time axis[9].

[9] The entropy is 0 when all respondents are characterized by the same category and is maximal when they are characterized by the same proportion of each category.

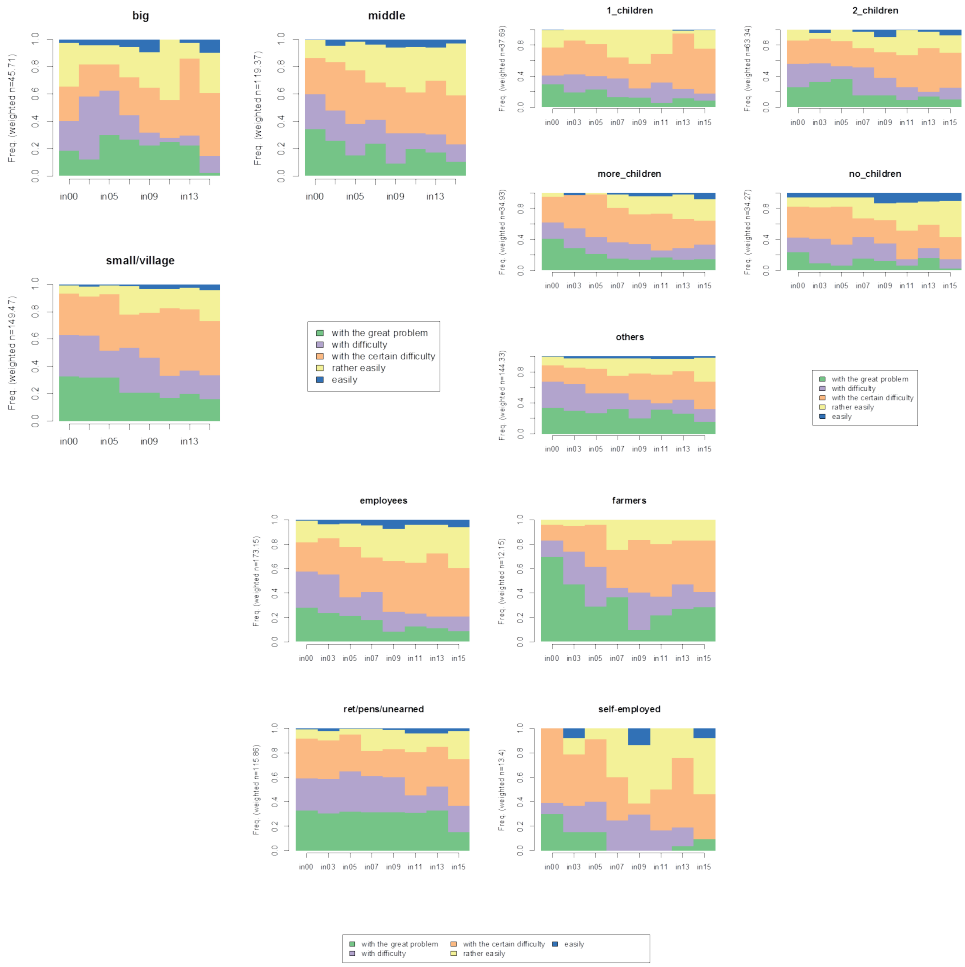

Fig. 5. Transversal categorical distributions by type of family, socio-economic group of household and place of living covariates

Source: own calculations in R.

Figure 6 shows the curves by family type, place of living and socio-economic group of households. For 'one child' family type the entropy index slightly decreases in 2013 and increases at the end of the follow-up period. This is a consequence of the increasing proportion of the third category of income assessment ("with certain difficulty") in 2013 and of improving financial situation as well as more balanced income levels distribution in 2015 (see Figure 5). Figure 6. shows almost a plateau at the level of entropy index for the retired, pensioners and living on unearned sources as opposed to the self-employed which can be explained by the significant jumps for this socio-economic group of households (see Figure 5).

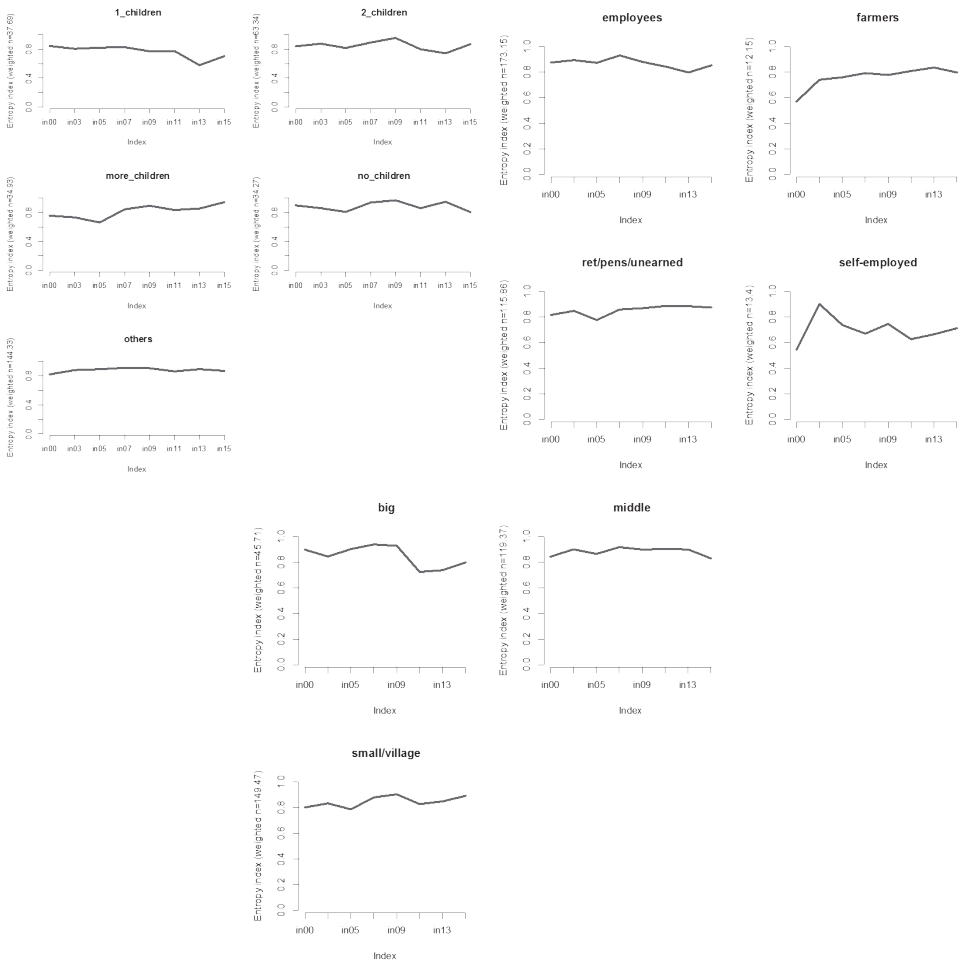

Fig. 6. Transversal entropy by type of family, socio-economic group of household and place of living covariates

Source: own calculations in R.

The highest value of the entropy index can be observed for households living in big cities especially in 2000-2009, then a considerable decrease in 2011 and 2013 could be noticed (this is a consequence of the increasing proportion of 'rather easily' and 'with certain difficulty' levels of income, respectively) followed by an increase of the entropy index at the end of the period.

5. Plotting individual sequence characteristics

Individual sequences can be characterized through:

a) Unidimensional indicators: *number of transitions in sequence v*: $l_d(v) - 1$, obtained from sequence length $l_d(v)$, *number of subsequences* $\varphi(v)$ and *within sequence entropy* [Gabadinho et al. 2011] given by:

$$h(p_1 \cdots p_c) = -\sum_{i=1}^{c} p_i \log p_i, \tag{1}$$

where c is the number of all categories and p_i is a proportion of occurrences of i-th category in the considered sequence.

b) Composite complexity measures: turbulence $T(v)$ [Elzinga, Liefbroer 2007] given by (2) and complexity $C(v)$ [Gabadinho et al. 2010] given by Equation (3):

$$T(v) = \log_2\left(\varphi(v)\frac{s^2_{t,\max}(v)+1}{s^2_t(v)+1}\right), \tag{2}$$

$$C(v) = \sqrt{\frac{l_d(v), h(v)}{l(v), h_{\max}}}, \tag{3}$$

where $s^2_t(v)$ is the variance of consecutive times for the response categories of sequence v. $s^2_{t,\max}(v)$ is the maximum value this variance can take given the total duration $l(v) = \sum_j t_j$ of the sequence. The maximum value is $s^2_{t,\max}(v) = (l_d(v) - 1)(1 - \bar{t}(v))^2$ where $\bar{t}(v)$ is the mean consecutive time spent in the different positions. $h_{\max} = \log c$ is the theoretical maximum value of the entropy. From the prediction point of view, the higher the differences in the durations of the category responses and hence the higher their variance, the less uncertain the sequence.

The minimum values for $C(v)$ equal to 0 can be achieved by the sequence represented by only one category (no transitions and $h(v) = 0$). The maximum value equal to 1 if i) $l(v)$ contains each of the c possible categories; ii) the same time is spent for each category $l(v)/c$; iii) there is the maximum number of transitions $l(v) - 1$.

Complexity measures take into account simultaneously for sequencing and durations.

In the theoretical left panel of Figure 7 we can observe the high turbulence value for the quite simple sequence [4] with a null variance of durations (this variance does not account for the positions that are not visited). The turbulence also exceeds the complexity index, i.e. for sequences [6] and [8] with a zero variance in duration.

The longitudinal entropy (that formulates the complexity index) discriminated clearly between the sequences with zero duration variance but does not account for the category order in the sequence (sequences [8] and [10] have the same maximal normalized entropy of 1, see Figure 7, left).

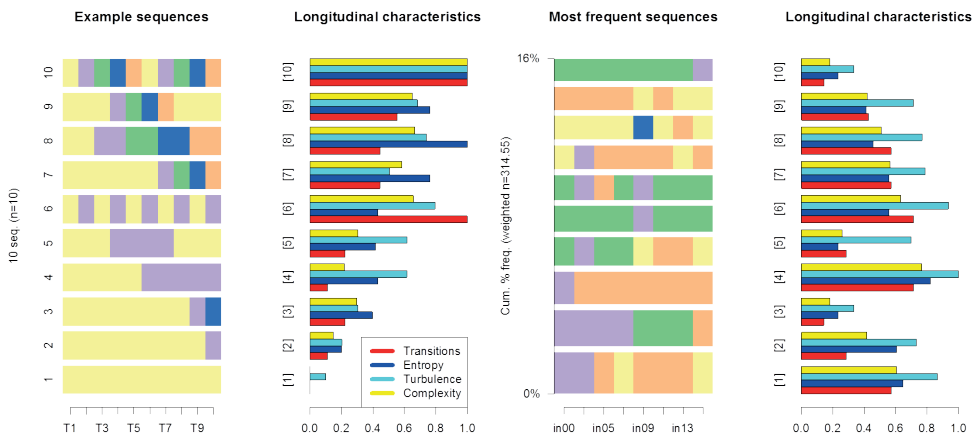

Fig. 7. Longitudinal characteristics for theoretical example (left) and income perception example (right)

Source: own calculations in R.

As far as the income perception is concerned (see Figure 7, right) the most frequent sequences are rather simple (not too many transitions are observed). There are only seven sequences (among all of them) with the highest number of transitions equal to 7. Moreover, the turbulence values exceed the complexity indices for all the presented sequences (with at least one response category which is not given). The maximum complexity value equal to 0.98 is achieved for two sequences GP-CD-D-RE-GP-D-CD-E and GP-E-D-CD-RE-CD-RE-E (not presented in Figure 7) showing also the improving financial situation in recent years.

6. Conclusions

We have presented the different graphical methods of the discrete longitudinal data analysis. On the basis of the presented figures we identified the demographic features of Polish families showing the improvement of the financial situation in recent years.

However, it should be noted that the sequences analysis is often very complicated to describe, visualize, and compare results (especially for large sequence data sets). Moreover, the results based mainly on the graphical presentations allow for unidimensional response variable analysis and assume the relationship of sequences with time-constant covariates (the covariates are not allowed to vary with time).

In future work we would like to examine the dynamics of the analyzed changes (in the financial situation of Polish families) using different latent variable models, especially the latent Markov models. These kinds of models are able to detect underlying latent structures and they can be used in various longitudinal settings: to account for measurement error, to detect unobservable states and also to include time-varying demographic features.

Bibliography

Billari F.C., 2001, *The analysis of early life courses: Complex description of the transition to adulthood*, Journal of Population Research, 18(2), pp. 119-142.

Brzinsky-Fay C., Kohler U., Luniak M., 2006, *Sequence analysis with stata*, The Stata Journal, 6(4), pp. 435-460.

Elzinga C.H., Liefbroer A.C., 2007, *De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis*, European Journal of Population, 23, 225-250, https://link.springer.com/content/pdf/10.1007%2Fs10680-007-9133-7.pdf.

Gabadinho A., Ritschard G., Müller N.S., Studer M., 2010, *Indice de complexité pour le tri et la comparison de séquences catégorielles*, Revue des nouvelles technlologies de l'information RNTI, E-19, pp. 61-66.

Gabadinho A., Ritschard G., Müller N.S., Studer M., 2011, *Analyzing and visualizing state sequences in R with TraMineR*, Journal of Statistical Software, 40(4), pp. 1-37.

Gabadinho A., Studer M., Müller N., Bürgin R., Fonta P.A., Ritschard G., 2017, *TraMineR – Trajectory Miner: A Toolbox for Exploring and Rendering Sequence*, Version 2.0-7, https://cran.r-project.org/web/packages/TraMineR/TraMineR.pdf.

Helske J., Helske S., 2017, *Hidden Markov Models for live sequences and other multivariate multichannel categorical time series*, Version 1.0.8, https://cran.r-project.org/web/packages/seqHMM/seqHMM.pdf.

Muthén B., Shedden K., 1999, *Finite mixture modeling with mixture outcomes using the EM algorithm*, Biometrics, *55*(2), pp. 463-469.

Ritschard G., Gabadinho A., Studer M., Müller N.S., 2009, *Converting between Various Sequence Representations*, [in:] Z. Ras, A. Dardzinska, *Advances in Data Management, Studies in Computational Intelligence*, 223 Springer-Verlag, Berlin, pp. 155-175, DOI:10.1007/978-3-642-02190-9_8.

Scherer S., 2001, *Early career patterns: a comparison of Great Britain and West Germany*, European Sociological Review, 17(2), pp. 119-144.

Singer J.D., Willett J.B., 2003, *Applied longitudinal data analysis: Modeling change and event occurrence,* Oxford, UK, Oxford University Press.

Social Diagnosis, 2015, *Objective and Subjective Quality of Life in Poland*, Czapinski J., Panek T. (eds.), Warszawa, Social Monitoring Council (22.11.2017), http://www.diagnoza.com/index-en.html.

Tueller S.J., 2017, l*ongCatEDA – Package for Plotting Categorical Longitudinal and Time-Series. Version 0.31*, https://cran.r-project.org/web/packages/longCatEDA/longCatEDA.pdf.

Tueller S.J., Dorn R.A., Bobashev G.V., 2016, *longCatEDA: Package for Plotting Categorical Longitudinal and Time-Series Data*, Methods Report RTI Press. 2016 Feb; 2016: MR-0033-1602, DOI: 10.3768/rtipress.2016.mr.0033.1602.

GRAFICZNE NARZĘDZIA PREZENTACJI DYSKRETNYCH ZBIORÓW PANELOWYCH W PROGRAMIE R

Streszczenie: Właściwa prezentacja graficzna danych jest przydatna podczas całego procesu analizy, począwszy od wstępnego rozpoznania zbioru do dopasowania modelu i prezentacji wyników. Najpopularniejszym sposobem wizualizacji danych panelowych jest oddzielne (dla każdej fali, w wymiarze przekrojowym) porównywanie wykresów dla każdego z okresów z osobna. Głównym celem tego artykułu jest przedstawienie zmieniającej się subiektywnej oceny sytuacji finansowej w Polsce dla wybranych cech demograficznych. W niniejszej pracy za pomocą różnych wykresów charakteryzujących sekwencje odpowiedzi (udzielanych przez respondentów w następujących po sobie okresach) przedstawione zostaną nowoczesne metody prezentacji dyskretnych zbiorów danych panelowych. Porównane zostaną również tzw. miary złożoności analizowanych sekwencji odpowiedzi. Obliczenia i wykresy przedstawione zostaną z wykorzystaniem głównie procedur pakietu `TraMineR` [Gabadinho et al. 2017] programu R.

Słowa kluczowe: dane panelowe, sekwencje odpowiedzi, wizualizacja danych.