



HAL
open science

SLIDER: Software for Longitudinal Data Exploration with R

Hadrien Commenges, Pierre Pistre, Robin Cura

► **To cite this version:**

Hadrien Commenges, Pierre Pistre, Robin Cura. SLIDER: Software for Longitudinal Data Exploration with R. *Cybergeo : Revue européenne de géographie / European journal of geography*, 2014, 693, 10p. 10.4000/cybergeo.26530 . hal-01213604

HAL Id: hal-01213604

<https://hal.science/hal-01213604>

Submitted on 12 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Cybergegeo : European Journal of Geography

GeOpenMod

Hadrien Commenges, Pierre Pistre et Robin Cura

SLIDER: Software for Longitudinal Data Exploration with R

Avertissement

Le contenu de ce site relève de la législation française sur la propriété intellectuelle et est la propriété exclusive de l'éditeur.

Les œuvres figurant sur ce site peuvent être consultées et reproduites sur un support papier ou numérique sous réserve qu'elles soient strictement réservées à un usage soit personnel, soit scientifique ou pédagogique excluant toute exploitation commerciale. La reproduction devra obligatoirement mentionner l'éditeur, le nom de la revue, l'auteur et la référence du document.

Toute autre reproduction est interdite sauf accord préalable de l'éditeur, en dehors des cas prévus par la législation en vigueur en France.

revues.org

Revues.org est un portail de revues en sciences humaines et sociales développé par le Cléo, Centre pour l'édition électronique ouverte (CNRS, EHESS, UP, UAPV).

Référence électronique

Hadrien Commenges, Pierre Pistre et Robin Cura, « SLIDER: Software for Longitudinal Data Exploration with R », *Cybergegeo : European Journal of Geography* [En ligne], GeOpenMod, document 693, mis en ligne le 07 novembre 2014, consulté le 07 novembre 2014. URL : <http://cybergegeo.revues.org/26530> ; DOI : 10.4000/cybergegeo.26530

Éditeur : CNRS-UMR Géographie-cités 8504

<http://cybergegeo.revues.org>

<http://www.revues.org>

Document accessible en ligne sur :

<http://cybergegeo.revues.org/26530>

Document généré automatiquement le 07 novembre 2014.

© CNRS-UMR Géographie-cités 8504

Hadrien Commenges, Pierre Pistre et Robin Cura

SLIDER: Software for Longitudinal Data Exploration with R

Introduction

- 1 In his classical handbook of data visualization, Tufte (1983) highlights the main characteristics a graphical display should fulfill. In particular it should "induce the viewer to think about the substance rather than about methodology, graphic design [...] or something else". It should also "encourage the eye to compare different pieces of data", "reveal the data at several levels of detail, from a broad overview to the fine structure" and "be closely integrated with the statistical and verbal descriptions of a data set" (p.13). These criteria are essentials but difficult to achieve, especially for time-oriented data. A brief glance at the more recent handbook of time-oriented data visualization (Aigner et al., 2011) is sufficient to show that a huge number of graphical displays does not satisfy Tufte's criteria, in particular the first one. Visual displays of time-oriented data often lack readability and force the viewer to focus on the methodology and the graphical design rather than on the substance.
- 2 First of all, we give an illustrated state of the art of existing graphical visualizations used to analyze longitudinal data, presenting step by step four graphical displays implemented in the well-known TraMineR R package (Gabadinho et al., 2011a): sequence distribution plot, sequence index plot, sequence frequency plot and parallel coordinates plot for sequence data. Then, we introduce the proposed "slide plot" visualization. It is a new graphical display which complements the existing ones and is particularly useful to identify dominant trends among individual trajectories of people or spaces. At last, the paper gives a technical description of a web application called SLIDER, implemented using the R software (R Core Team, 2013) and the shiny R package (RStudio Inc. 2014). This interactive platform is conceived for exploring longitudinal, it includes the TraMineR plots and the proposed *slide plot*.

Visualization of longitudinal data: state of the art

- 3 This section introduces the main graphical displays for exploring longitudinal data, highlighting their benefits and their drawbacks. Each graphical display is provided with two empirical examples and a short thematic comment. The graphical displays are produced with the R software and the TraMineR package, implemented by Gabadinho et al. (2011a, 2011b).

Datasets

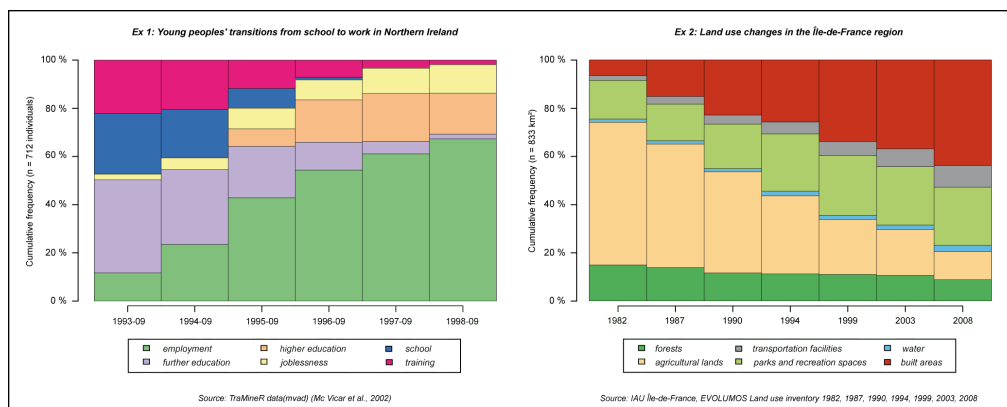
- 4 Two datasets are used to exemplify this state of the art: the first one is taken from Mc Vicar's paper about young peoples' transitions from school to work (Mc Vicar et al., 2002). The data comes from a cohort survey implemented in Northern Ireland which follows a sample of 712 young individuals aged 16 in 1993. Monthly labor activity is collected during six years, from 1993 to 1998. Six activities are coded: school, further education, higher education, employment, joblessness and training. In this example, the statistical units are persons and the variables are categories of labor activity declared at each time step (we only take into account the activity in September of each year).
- 5 The second dataset is the land use basemap produced by the planning institute of Île-de-France (IAU-IdF - *Institut d'Aménagement et d'Urbanisme d'Île-de-France*). The data consists of a time series of geolocated polygons describing the land use changes from 1982 to 2008 for the whole Île-de-France region (IAU-IdF, 2013). Six aggregated land use categories are coded from the original typology of 80 categories: forests, agricultural lands, water, parks and recreation spaces, transportation facilities, built areas. In this example, the statistical units are portions of land and the variables are categories of land use at each time step. Two precisions are needed to understand this example. First, we work on a selection of the portions of land which have experienced a change of land use between 1982 and 2008. This selection represents

7% of the whole area: 833 km² for a total area of 11,200 km². Then, each geolocated polygon is weighted by its area. The weighting variable affects all the numeric and graphic outputs.

Review of the existing graphical displays

- 6 Two main graphical displays are used in the literature: the sequence distribution plot and the sequence index plot. These displays are mainly used to visualize trajectories of individuals: professional trajectories, family trajectories or scheduling¹.
- 7 The sequence distribution plot shows the distribution of statistical units within the defined categories at each time step, *i.e.* a cross-sectional frequency. The sequence index plot preserves the individuality of trajectories, which may be useful to visualize a sample of the studied population. Each sequence represents one statistical unit as a set of colored segments corresponding to its successive states. The oldest sequence index plot we found for this state of art was proposed by Bonnafous et al. (1981). It consists of a set of cards with colored edges: each card represents an individual and the colored segments on its edge show the trips and the activities carried out during the day.
- 8 The following graphical displays (Figure 1, Figure 2, Figure 3, Figure 4) illustrate the main existing possibilities to visualize longitudinal data. Each display is applied to both datasets – young persons' professional trajectories in Northern Ireland and changes of land use in the Île-de-France region.
- 9 **Display 1** – The first example is the sequence distribution plot (Figure 1) which shows, at each time step, the cross-sectional frequencies within the defined categories. This display highlights the main categories at each time step and the evolution of their respective importance during the studied period.

Figure 1: Two examples of sequence distribution plot



Ex.1 – Young people's professional trajectories: at the beginning of the period, the most frequent categories are the schooling ones: school and further education. The employment situation progressively grows in importance and becomes dominant in 1998.

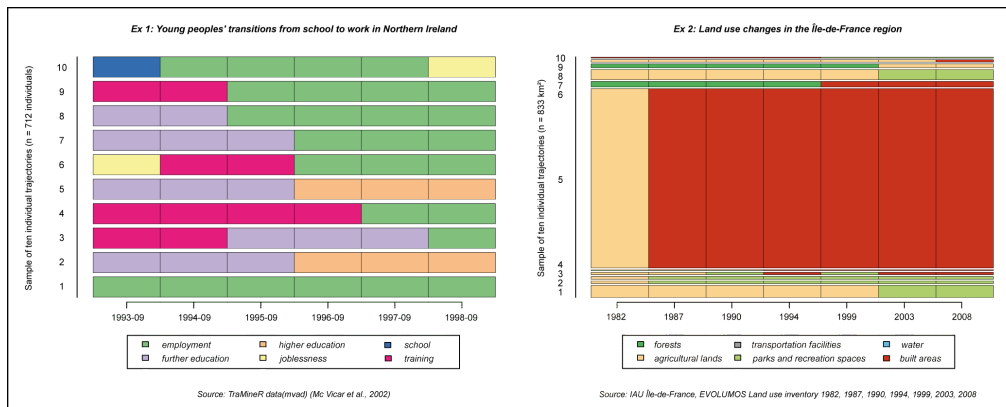
Ex.2 – Changes of land use: among the portions of land that experienced a change of land use between 1982 and 2008, the agricultural land is the main category at the beginning of the period but it decreases through time. To a lesser extent, the forest land use follows the same trajectory. On the other hand, the urbanized land (urban open spaces, built areas, transportation facilities) grows in importance during the period.

Drawbacks: the graphical display doesn't show the successive steps experienced by the statistical units (individuals or portions of space) to move from one state to another. For example, we can't know if the growing area of the transportation facilities is provided by the decreasing area of agricultural land use, by the decreasing area of the forest land use, by both or by other growing land uses. Moreover, this display isn't strictly longitudinal: it could be applied to a cross-sectional sample of the general population at each time step.

- 10 **Display 2** – The second example is the sequence index plot (Figure 2). It represents a sample of the first ten persons or polygons' trajectories of each dataset, describing each state at each step of each individual. For Ex.1, each trajectory is composed of one (individual n°1), two (individuals n°2, n°4, n°5, n°7, n°8, n°9) or three (individuals n°3, n°6, n°10) successive states.

For Ex.2, all individual trajectories are split in two states except for individual n°4 (three states).

Figure 2: Two examples of sequence index plot



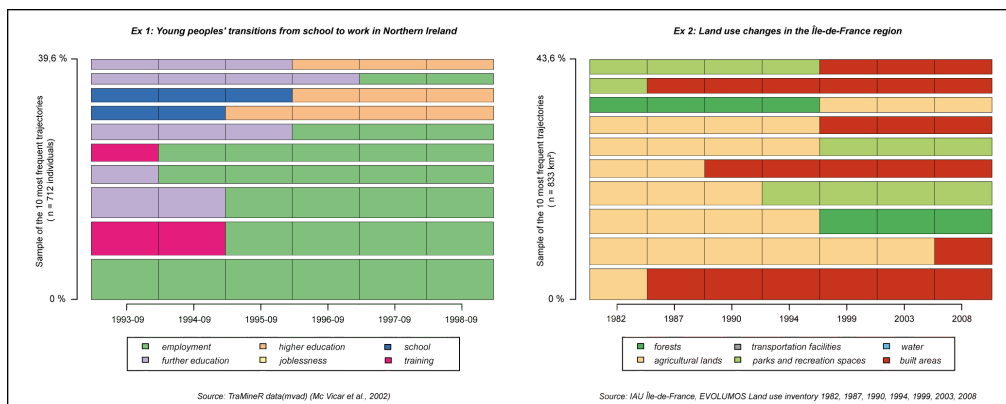
Ex.1 – Young people’s professional trajectories: comparing the ten sampled individual trajectories reveals some recurrences: strong representation of further education and training states at the beginning, further education and employment states in the middle, and employment at the end of the period, but also a majority of transition from further education or training states to employment state during the whole period.

Ex.2 – Changes of land use: most of the ten sampled trajectories (n°1, n°2, n°3, n°4, n°5, n°7, n°9) describe land use changes from agricultural lands to parks and recreational spaces or built areas. The width of the segments shows the variability of surface area, with the individual n°5 standing out from the rest of trajectories.

Drawbacks: the sequence index plot is a strictly longitudinal display because it represents the succession of states for each individual. Nevertheless, the sample of individuals isn’t representative of the whole population. This graphical display can also be used to represent the whole dataset but it lacks readability because of over-plotting.

11 **Display 3** – In line with the second example, the third one is the sequence frequency plot (Figure 3) which also represents individual configurations of trajectories by graphical overlay. But in this case, it shows the ten most frequent trajectories of each dataset. This visualization is a form of display summary of the previous two examples, both keeping the value of an aggregate analysis of the whole sample and a true representation of individual trajectories.

Figure 3: Two examples of sequence frequency plot



Ex.1 – Young people’s professional trajectories: the ten most frequent trajectories account for 39,6 % of the whole studied population (712 individuals). The main profile trajectory is a unique employment state from September 1993 to September 1998. The six following main frequent configurations begin by training or further education states and continue by employment periods.

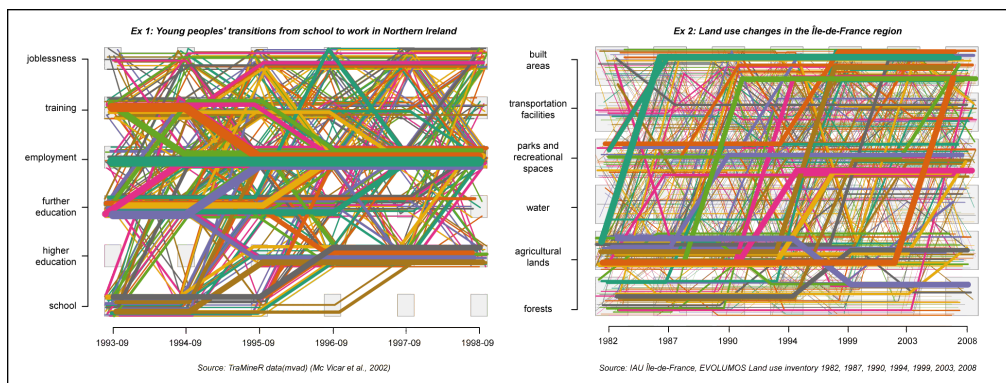
Ex.2 – Changes of land use: the ten most frequent trajectories account for 43,6 % of the whole studied population (180,000 polygons weighted by their area). The main one shows land use

changes from agricultural lands to built areas between 1982 and 1987. The six following main frequent trajectories also report decreasing of agricultural lands in favor of built areas, forests or parks and recreation spaces.

Drawbacks: the sequence frequency plot is a good combination of aggregated and individual aspects of the trajectories: For this purpose, only the exact same trajectories are aggregated. That is the reason why the display is useless to detect the most frequent changes from one state to another, neither the intensity of these changes nor their temporality.

12 **Display 4** – The last example is the parallel coordinates plot for sequence data, recently developed by the TraMineR package team (Figure 4) (Bürgin et al., 2012). This visualization is inspired by the classical parallel coordinates display. It aims to show the most frequent patterns aggregating the same trajectories or part of trajectories, while maintaining the individual aspect and representing the changes of states. It can be used both to represent trajectories between successive states (one by each time step) or events (more than one simultaneous events by each time step).

Figure 4: Two examples of parallel coordinates plot for sequence data



Ex.1 – Young people’s professional trajectories: the parallel coordinates plot for sequence data reveals concentrations of individuals for some states at certain time steps (eg: further education and school at the beginning or higher education and employment at the end). It also highlights the most frequent trajectories (eg: the only employment state from 1993-09 to 1998-09 or the training state to 1994-09 following by a whole period of employment), but they are difficult to identify and to rank precisely.

Ex.2 – Changes of land use: agricultural lands state in 1982 and built areas state in 2008 are the two main states which concentrate most of the main trajectories beginnings and ends. Furthermore, the most frequent trajectories show land use changes from agricultural lands to built areas, parks and recreational spaces or forests, but it is also difficult to identify in which periods the changes were the most numerous.

Drawbacks: the parallel coordinate plot for sequence data is not suitable to represent large samples of longitudinal data, especially when the individual trajectories are complex (several states or events and significant heterogeneity in individual behavior).

13 To conclude this short state of the art, the sequence distribution plot (Figure 1) provides a cross-sectional view of the distribution of states at each time step, but it loses the longitudinal aspect of the dataset. The sequence index plot (Figure 2) is readable only for a small sample which isn’t representative of the whole population. The sequence frequency plot (Figure 3) fails to detect the most frequent changes from one state to another, neither the intensity of these changes nor their temporality. Finally, the parallel coordinates plot for sequence data (Figure 4) provides graphical solutions to these drawbacks, but the conservation of the individual aspect of the trajectories causes a loss of readability of the main trends that emanate from all individual trajectories.

The slide plot

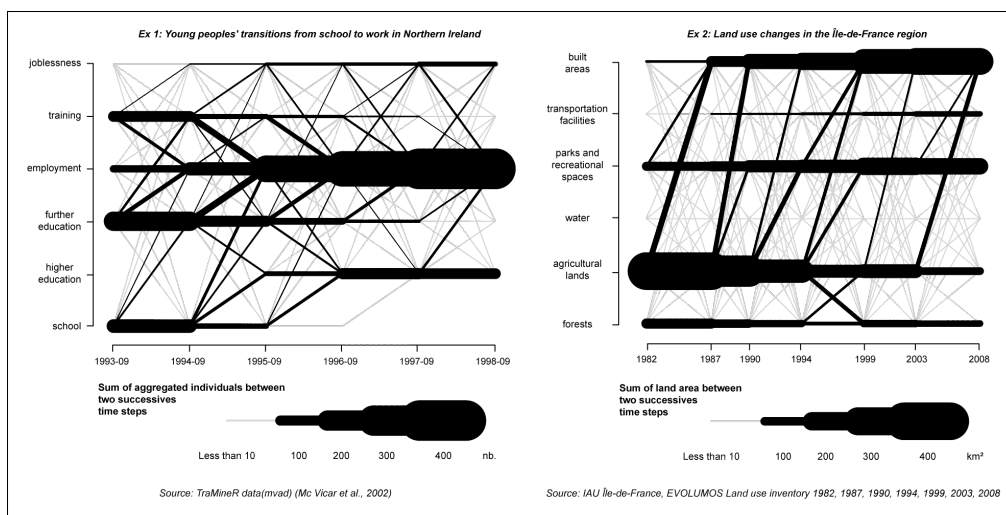
14 This short state of the art shows several existing graphical displays. Each of them is useful to visualize different aspects of a phenomenon grasped by longitudinal data. The so-called slide

plot is proposed to fill an important limitation of the existing graphical displays: they fail to efficiently represent aggregated tendencies.

The slide plot: examples of application

- 15 The slide plot consists in aggregating the common parts of the individual trajectories (Figure 5). This aggregation aims to show the dominant trends by hiding the high variability of the individual trajectories. The slide plot can be interpreted through two readings: the first one consists in bringing out the main trends for the whole sample; the second one consists in detecting specific relationships between two or more categories.
- 16 The graphical display is built as a cross-classification of two categorical variables: on the X-axis goes a discrete time variable (time steps); on the Y-axis goes a nominal or ordinal variable corresponding to the modalities of a categorical variable (states). The core of the plot is made up of a set of segments. The visual variation of the segments' width represents the absolute frequencies of the aggregated individual trajectories. This frequency may be weighted by a weighting variable, as is the case for the land use example.

Figure 5: Two examples of slide plots



Ex.1 – Young people's professional trajectories: the slide plot brings new elements of interpretation: school, training and further education are indeed the most frequent categories at the beginning of the period. Then, several trends emerge: (1) some of the students continue studying (higher education) immediately after school or after a transitional period of further education; (2) a high number of students enter in the labor market, immediately after school but mainly after a period of further education or training. In conclusion, the slide plot brings out the main trends: continuation of the education phase and entrance in the labor market.

Ex.2 – Changes of land use: the previous graphical displays showed that, among the portions of land that experienced a change of land use between 1982 and 2008, the agricultural land is the main category at the beginning of the period and the built areas is the main category at the end. The slide plot shows the relationships that produce this result. Urbanized areas are built on forest and agricultural lands, but not exclusively: there are significant relationships between growing categories (from urban open space to build areas) and relationships between decreasing categories (reciprocal changes between forest and agricultural land).

Drawbacks: As with any graphical display, a difficult trade-off must be made between the preservation of information and the graphical readability. The slide plot brings out the dominant trends at the expense of the individual aspect of trajectories.

The slide plot: technical implementation

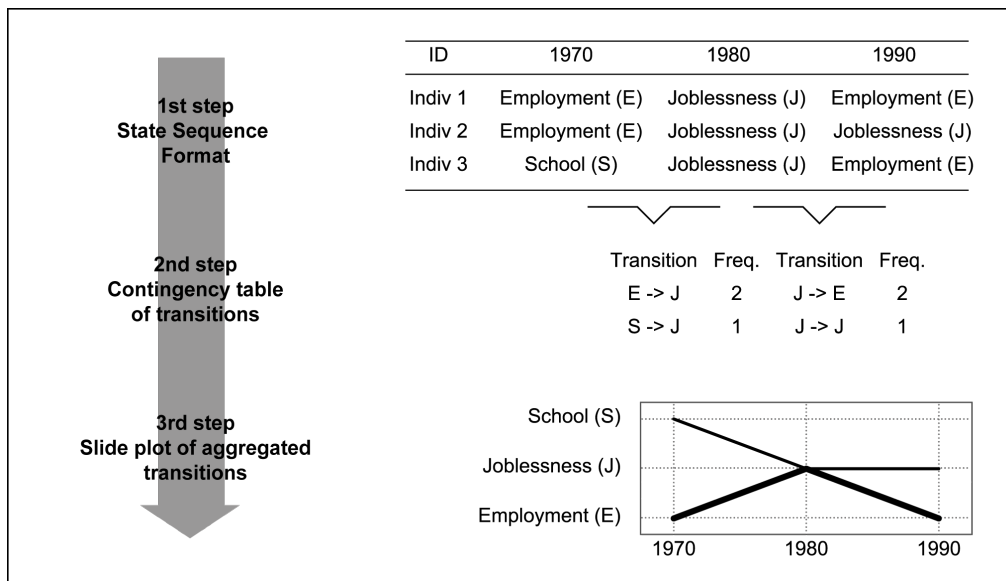
- 17 The slide plot function is available on the following GitHub repository: <https://github.com/hcommenges/slider>. It relies on two R packages: *reshape2* (Wickham, 2007) and *ggplot2* (Wickham, 2009). The first one is used to reshape the data, the second one implements the grammar of graphics (Wilkinson, 2005) used here to design the slide plot.

18 The input data format is the so-called STate-Sequence (STS) format that is already used by TraMineR functions (Gabadinho et al., 2011b). Each individual is described by cross-classifying two categorical variables: a time variable corresponding to successive time steps and a factor variable (nominal or ordinal) characterizing the statistical units (sex, profession, age groups, etc.). The input may also include a weighting variable. This option is useful if the data is a sample coming from a survey with weights; it is also useful to weight the statistical units by a variable of interest, such as the surface area as is the case of the land use dataset presented below.

19 In the following example (Figure 6), we have three individuals described by three time variables (1970, 1980, 1990). These time variables characterize the state of the individual at each time step. Considering the whole set of states, we define the alphabet as the list of distinct modalities found in the time variables. In this example, the alphabet is made of three distinct states: School, Employment and Joblessness.

20 From the state sequence format, a contingency table of transitions is computed for each couple of successive time steps. This contingency table is then plotted through a set of articulated segments. Their thickness is proportional to the absolute frequency of transitions in the simplest case. It is proportional to the weighted frequency if there is a weighting variable.

Figure 6: From the input table to the slide plot output

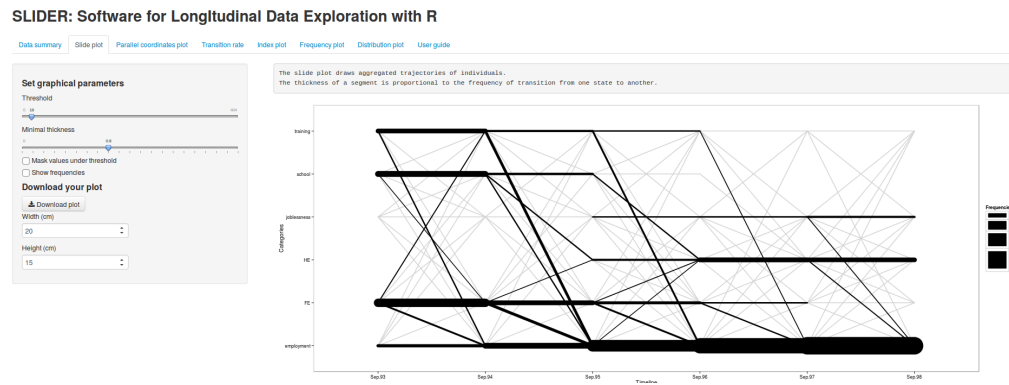


21 The main difference between the slide plot and the parallel coordinates plot implemented in the TraMineR package (Bürgin et al., 2012) is that we consider that all the sections of the trajectories are embeddable. By losing the individual aspect of trajectories we increase the readability of aggregated tendencies.

The SLIDER application

22 The SLIDER application is available at <http://slider.parisgeo.cnrs.fr> and its code is available on the following GitHub repository <https://github.com/hcommenges/slider>. It is conceived as an interactive tool for graphical exploration of longitudinal data. It is built with the shiny R package (RStudio Inc., 2013), which provides a concise and efficient way to build and publish a web application using R code.

23 SLIDER is composed of a side panel and a main panel (Figure 7). The side panel provides options to load the data and settings for the interactive selections. The main panel shows seven tabs providing tables and plots to explore longitudinal data: data summary, transition rate, sequence distribution plot, sequence index plot, sequence frequency plot, parallel coordinates plot for sequence data and slide plot.

Figure 7: Screenshot of the SLIDER application

24 The side panel proposes several buttons and boxes:

- **Load example data.** This button loads the data from McVicar and Anyadike-Danes (2002) included in the TraMineR package.
- **Upload CSV file.** The user may load its own data saved in a text file. The default parameters are comma as separator and double quote as quoting character. Those settings can be changed, before or after uploading the CSV file, by checking the CSV options checkbox.
- **Choose time steps.** Two or more fields containing the time steps may be chosen to explore the dataset.
- **Choose weighting variable.** One field containing the weights assigned to the individuals.
- **Choose factors.** A field containing a qualitative variable may be chosen to explore the patterns for different groups (e.g. the sex of the individuals).
- **Choose a group.** The value assigned to the factor variable group (e.g. "male" or "female" if the chosen factor is the sex of the individuals) for which you want to explore the data.

25 You can perform interactive selections by choosing up to two factor variables and, for each of these, one or several modalities. The performed selection is an intersection (AND boolean operator): for example, by choosing "female" (sex) and "retired people" (occupational status), the pseudo-syntax query is as follows: WHERE SEX = "FEMALE" AND STATUS = "RETIRED". The output of this selection is a sub-population of retired women.

26 The main panel is made of seven tabs providing tables and plots to explore longitudinal data. Each of these outputs can be downloaded with the Download button. It renders a SVG vector file with a default size of 20cm (w) x 15 cm (h). You can set the size from 1 to 30 cm. Be aware that index plots can be disk-space consuming when involving a lot of time steps and/or trajectories.

27 The **data summary** tab indicates the number of observations, the number of variables and shows a sample of 25 rows of the table. The default number of rows may be changed and the table can be searched, filtered and re-ordered.

The **transition rate** tab shows the frequency of transitions from one state to another, as observed in the dataset. Three options are available: absolute frequencies, row percentage and column percentage.

28 The **slide plot** tab draws the graphical display described above. Four options are available:

- **Threshold.** The slide plot might draw a huge number of micro variations which may harm the readability. Values under the set threshold are treated as residuals. The threshold value is considered as the minimal value.
- **Minimal thickness.** The thickness of each segment is proportional to the frequency of the aggregated trajectories it represents. The thickness depends on three parameters: the threshold parameter, the maximal value of the dataset and the minimal thickness parameter. The minimal thickness is assigned to the threshold value and the maximal thickness is assigned to the maximal value. For that reason, you can modify the overall appearance using both parameters: threshold and minimal thickness.

- **Mask.** After setting the threshold, you can choose to draw or to mask the micro variations. If you mask them, they won't appear, if you draw them, they will appear in light grey.
- **Frequencies.** This option draws the frequency of transition for each segment above the threshold.

29 Finally, the SLIDER application proposes four tabs to draw the TraMineR plots (Gabadinho et al., 2011a, 2011b): parallel coordinates plot for sequence data, sequence index plot, sequence frequency plot and sequence distribution plot. For these plots, the SLIDER application provides some options described in the TraMineR package documentation.

Conclusion

30 The SLIDER application provides a user-friendly interactive platform to explore longitudinal data and to export graphical results in vector files. It offers a new graphical display, the "slide plot", which is useful to explore a wide range of phenomena grasped by longitudinal. It provides a graphical interface to render the slide plot (SLIDER application) and it provides the R code to load the slide plot function in a native R program (SLIDER code). In addition, the SLIDER application grants access, in a user-friendly interface, to four graphical displays implemented in the TraMineR package.

Bibliographie

Aigner W., Miksch S., Schumann H., Tominski C., 2011, *Visualization of time-oriented data*, London, Springer.

Bonnafous A., Patier-Marque D., Plassard F., 1981, *Mobilité et vie quotidienne*, Lyon, Presses universitaires de Lyon.

Bürgin R., Ritschard G., Rousseaux E., 2012, "Exploration graphique de données séquentielles" in *Actes de l'atelier Fouille Visuelle de Données : méthodologie et évaluation - conférence Extraction et Gestion de Connaissances*, Bordeaux, 31 janvier.

Gabadinho A., Ritschard G., Mueller N.S., Studer M., 2011a, "Analyzing and Visualizing State Sequences in R with TraMineR", *Journal of Statistical Software*, vol. 40, No.4, 1-37.

Gabadinho A., Ritschard G., Studer M., Müller N. S., 2011b, *Mining sequence data in R with the TraMineR package: A user's guide*, University of Genève, <http://mephisto.unige.ch/traminer/index.html>.

IAU-IdF, 2013, *MOS 1982-2012. Du ciel à la carte*, Les Cahiers, Paris, Institut d'Aménagement et d'Urbanisme d'Île-de-France, No.168.

McVicar D., Anyadike-Danes M., 2002, "Predicting successful and unsuccessful transitions from school to work by using sequence methods", *Journal of the Royal Statistical Society, Series A*, vol.165, No.2, 317-334.

R Core Team, 2013, *R: a language and environment for statistical computing*, Vienna, R Foundation for Statistical Computing, <http://www.R-project.org/>.

RStudio Inc., 2013, *shiny: Web Application Framework for R*, <http://CRAN.R-project.org/package=shiny>.

Tufte E.R., 1983, *The visual display of quantitative information*, Cheshire, Graphics Press.

Wickham H., 2009, *ggplot2: elegant graphics for data analysis*, New York, Springer.

Wickham H., 2007, "Reshaping Data with the reshape package", *Journal of Statistical Software*, vol.21, No.12, 1-20

Wilkinson L., 2005, *The grammar of graphics*, 2nd edition, New York, Springer.

Notes

1 See the references on the website of the TraMineR package: <http://mephisto.unige.ch/traminer/user.shtml>

Pour citer cet article

Référence électronique

Hadrien Commenges, Pierre Pistre et Robin Cura, « SLIDER: Software for Longitudinal Data Exploration with R », *Cybergeo : European Journal of Geography* [En ligne], GeOpenMod, document 693, mis en ligne le 07 novembre 2014, consulté le 07 novembre 2014. URL : <http://cybergeo.revues.org/26530> ; DOI : 10.4000/cybergeo.26530

À propos des auteurs

Hadrien Commenges

UMR 8504 Géographie-cités
13, rue du Four, 75006 Paris
hcommenges@parisgeo.cnrs.fr

Pierre Pistre

IFSTTAR, AME, DEST
14-20, boulevard Newton, Cité Descartes, Champs sur Marne
77447 Marne la Vallée Cedex 2
pierre.pistre@ifsttar.fr

Robin Cura

UMR 8504 Géographie-cités
13, rue du Four, 75006 Paris
robin.cura@parisgeo.cnrs.fr

Droits d'auteur

© CNRS-UMR Géographie-cités 8504

Résumés

This paper introduces an interactive web platform called “SLIDER” to explore longitudinal data and an original graphical display called “slide plot” which is conceived to visualize aggregated trajectories. The paper begins with a short state of the art of existing graphical displays used to analyze longitudinal data. Then, it presents the main characteristics of the proposed slide plot visualization. At last, it gives a technical description of the web application and the graphical display, both implemented using the R software and the shiny R package.

Cet article présente une plateforme web interactive baptisée “SLIDER” et un type de graphique original baptisé “graphique en coulées” (slide plot), ces deux outils étant conçus pour explorer des données longitudinales. L'article commence par un court état de l'art des modes de visualisation existants pour analyser les données longitudinales. Il poursuit par une présentation de l'usage et des caractéristiques techniques du graphique en coulées. Enfin, il décrit la plateforme interactive mise en place avec le *package* shiny du logiciel R.

Entrées d'index

Mots-clés : visualisation interactive, données longitudinales, logiciel R, application shiny

Keywords : interactive visualization, longitudinal data, R software, shiny web application