Analysis and Visualization of Austria's Social Insurance System

Erich Heil
Institute for Advanced Studies
Josefstädterstraße 39
Vienna, Austria
heil@ihs.ac.at

Dominik Walch
Institute for Advanced Studies
Josefstädterstraße 39
Vienna, Austria
walch@ihs.ac.at

ABSTRACT
We analyze the Austrian social insurance system, focusing on the labor market as well as the pension system, on behalf of governmental institutions and other official stake holders. Visualizations play an increasingly important role. We present two examples from current projects – a very simply static one and a more complex interactive one – that both demonstrate the benefits of visualizations especially in communicating results and insights. We aim to outline the main steps of our workflow beginning with data transformation up to interactive visualizations.

Categories and Subject Descriptors
A.0 **[General Literature]**: General Literature – *Conference proceedings;*
H.2.4 **[Database Management]**: Systems – *Relational Databases;*
H.5.0 **[Information Interfaces and Presentation (e.g., HCI)]**: General

General Terms
Algorithms, Economics

Keywords
Austrian social insurance data, data, data structures, interactive visualizations, chord diagram, data transformation

1. INTRODUCTION
At the Institute of Advanced Studies (IHS) we regularly analyze various aspects of the Austrian labor market, the health system and the pension system on behalf of Austrian ministries and other official stakeholders[1][2][3].

We increasingly rely on administrative micro data (mostly from the Austrian social insurance system) for these analyses. Replacing highly aggregated (macro) data allows us asking and answering new questions, and shedding light on phenomena so far undiscovered.

However, these increased possibilities come at a price: the cost of adapting the mindset, the tools and the interdisciplinary approach of modern data science.

The size and the complexity of the data (which we will briefly describe in section 2) call for new methods of analysis and new approaches in communicating the results, especially to a non-technical audience.

Visualizations proved to be one method that facilitates the analysis of the data as well as the presentation of results.

In section 3 we will present two examples of visualizations - a simple static one and a more complex interactive one. Both examples are drawn from ongoing projects. In both examples using visualizations was successful from our practitioners' point of view. In section 4 we will conclude with discussion and will give a brief outlook on future research in section 5.

2. DATA
Austria`s social insurance system encompasses the health system, the pension system and the (un)employment system and covers roughly 99 percent of Austria's population. The Austrian Social Insurance legislation defines over 300 different states that describe a person's situation at any given point in time (multiple states are possible). This includes all types of employment status as well as a non-employment status, such as maternity leave, army service, retirement, and so on.

It is a legal requirement to document the different social insurance status of a person. The precise status has to be electronically recorded, including starting date, ending date, the firm (or in case of a non-employment state such as unemployment, the relevant institution) and the wage received. All the information is electronically recorded and stored in a database.

Scientific institutions such as IHS have access to an anonymized version of the database run by the Ministry of Social Affairs (BMASK) and the public unemployment agency (AMS).

Technically we deal with a relational database, persons and firms being the entities and the social insurances spells[1] being the relations. Every entity has a unique identifier. We have personal characteristics such as age, gender, nationality and firm features such as the region and the sector of economic activity[2].

3. Two examples
We will present two examples where visualizations played a crucial role. Both examples deal with employment stocks and flows. They are in the vain of the literature on labor market flows as they analyze the creation and destruction of jobs, and the flow of workers [4].

[1] Spells are relations, which describe the match between employers and employees in a period of time. Example: I (employee) am employed at Billa (employer) from 1.1.2013 to 2.3.2014 (period of time).

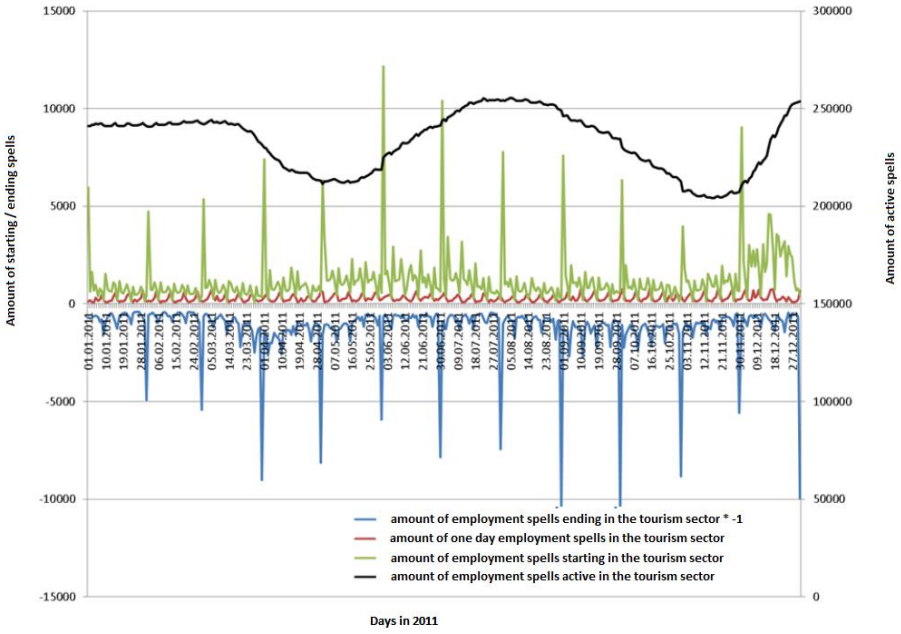
[2] Coded as NACE "Nomenclature statistique des activités économiques dans la Communauté européenne", example : I – Tourism.

The first example is an analysis of employment stocks. By using daily instead of monthly data and augmenting the graphical representation with the display of the number of jobs that start (are created) and end (are destroyed) on each particular day, we could provide insight during the legislation process surrounding the introduction of the current lay-off tax (Auflösungsabgabe)[3].

The second example deals with flows of workers. We analyze intra- and inter-sectoral flows of workers using an interactive chord diagram[4]. We will show the necessary data preparation steps as well as the interactive visualization and examples of our findings.

3.1 Simple Example with Policy Impact

Figure 1 illustrates the benefit of more detailed data even for very simple analyses. We display the stock (sum of all active spells per day) of employment in the tourism sector (2011) on a daily basis[5] (black line- right axis). In addition, we also display the number of employment spells that start and those that end on a particular day (green and blue line – left axis).



Inspecting the black line – the daily stock of employees - we can see – firstly, the well-known and -documented phenomenon of seasonal employment peaks during summer and winter season and secondly, exactly 52 small bumps representing employment peaks during weekends.

Looking at the green and blue line (the red line represents episodes of one-day length), we can see that jobs are being created on Fridays and Saturdays and terminated (destroyed) on Sundays.

Additionally we see that, in tourism, jobs are permanently created and destroyed which is consistent with the high employee-turnover of the sector. Even in times of diminishing employment – stock new jobs are created and conversely even in times of increasing employment, jobs end. Burgess [5] for instance, investigated such phenomena.

In early 2012 the Austrian government started the legal process for the introduction of a lay-off tax. Whenever an employment episode is terminated a certain fee (€ 118 in 2015) is to be paid as a tax or "fine". The proceeds from the tax are to finance projects of the unemployment agency (AMS). The government was not aware of the fact that in certain sectors especially in tourism short term (peak) jobs are created and destroyed at a very high frequency. Additional labor costs of 118 € are negligible in longer lasting jobs. In the case of one, two or three days jobs such a layoff-tax would have driven a too large wedge between wage and productivity.

Using figure 1, we consulted the Ministry of Economic Affairs and the Chamber of Commerce during the early months of 2012. The proposed legal text was adjusted, very short term employment spells and internships were exempt from the tax.[6]

3.2 Dynamic Aspects – Inter and Intraindustry Worker Flows

In the generic case: whenever a job ends a worker enters into a state of non-employment. Sometimes later, a new job starts and the worker enters a new employment spell. Bjelland [6] for instance analyze such transitions from one employer to another.

As the Austrian social insurance data provides unique identifiers for persons and firms we can track people and observe their employment spells at their different employers over the course of time. This allows us to explicitly construct the transitions (flows) of workers from a source to a target employment.

In our example we are interested in flows of workers within and between sectors of economic activity, i.e. intra and inter-sectoral flows. To answer question like:

[3] https://www.ris.bka.gv.at/Dokument.wxe?Abfrage=Bundesnormen&Dokumentnummer=NOR40138051

[4] "Graphical method for displaying the inter-relationships between data in a matrix. The data is arranged radially around a circle with the relationships between the points typically drawn as arcs connecting the data together." From: https://en.wikipedia.org/wiki/Chord_diagram

[5] this differs from the standard publications of Hauptverband der Sozialversicherungsträger (HV) that use 12 data points describing the stock at the end of each month

[6] https://www.ihs.ac.at/publications/lib/ots_01242012.pdf

- Which economic sectors have high internal flows?
- Which sectors are connected by workers flows, i.e. where do sectors recruit their (new) personnel from?

As a first step we transform the data and construct the flows. In a second step we extract appropriately aggregated flows and visualize them in an interactive manner.

3.2.1 Data Transformation

To create a flow-table, (i.e. a table containing source and target employment spells as well as an explicit description of the transition) we take the following steps[7]:

1. Take all employment spells and save it to a table (T)
2. For each spell in T look at the next day and find the spells in a copy of T (R) which start or are active for the same person. Save the match consisting of spells in T (source) and R (target) in a table (Y). Delete from T all spells in source of Y.
3. For each spell remaining in T find the spell/s in R which start later then the end of spell in T +1 day for the same person. Only take the spells in R which have the earliest starting date. Save the match consisting of spells in R (source) and Y (target) in a table (X). Delete from R all spells in source of X.
4. For all remaining spells in R create a new table Z consisting of spell of R as source and fictional spell as target (pointing to the future).
5. Redo point 1 to 4 but instead of looking into the future (end of a spell) look into the past (start of a spell). This creates tables Y2,X2,Z2.
6. Union all flows (source and target) of Y,X,Z,Y2,X2,Z2 and eliminate all duplicates. Save this as final flow table.

Our employment spells are stored in a PostgreSQL table (60 Million rows) and we implement the algorithm in SQL to create the final flow table. Having this table we can filter for time intervals (like year), source information (like sector of economic activity) or region (e.g. "Bundesland") target information, attributes of the person involved (sex, age, ..), and properties of the transition itself[8].

3.2.2 Interactive Visualization of Inter- and Intra Industry Flows – using a Chord Diagram

As a template [8] is freely available and as it seemed both intuitive and aesthetically pleasing, we have chosen to visualize the intra and inter-sectoral worker flows in chord diagrams.

Using a Python script, we

- extract a transition matrix from the flow-table
- create an adapted version of the Global Migration Data Sheet [8] that allows us to
- generate interactive chord diagrams.

We create different chord diagrams for each year, for different regions ("Bundesland") and for different types of the flows. In each chord diagram one can further filter by the personal characteristics gender and nationality[9].

Figure 2 illustrates all inter and intra-sectoral flows for the year 2011.

3.2.3 Insights from the Visualizations

Exploring the different aspects of the visualizations and confronting them with our domain knowledge we could generate insights and confirm prior knowledge about the Austrian labor Market:

Comparing 1999 with 2011 we see an increase in the importance of the sector ÖNACE N – Administrative and Support Service: In 2011, this sector had one of the largest shares in terms of total in- and outflows as well as in terms of intra-flows. This is consistent with the relative growth of the tertiary sector and the fact that ÖNACE N – includes firms in the sector N 78: Employment Activities. Especially the more volatile jobs of several sectors are transferred to such firms.

When inspecting secondary employment we see (i) large inflows from self-employment i.e. people who are self-employed often take up interim employment in various other sectors and (ii) ÖNACE I - Tourisms is by far the most important sector for (short) secondary employment. This is consistent with our employment peaks during weekends.

When comparing employees of different nationalities we see different flow-patterns that are consistent with differing industry attachments of these groups as well as the entry (and exit) into (from) the Austrian labor market. In 2011 the largest inflow into the Austrian labor market originated in the surrounding "EU-Neu"[10] countries.

[7] The algorithm was constructed in a different ongoing project [7] where we create a stock-flow-consistent data framework that explicitly models flows as pairwise links between two subsequent employment spells. The framework is general as it does not only cover the generic case, but allows for multiple employment spells, secondary employment and overlapping spells.

[8] These include, for instance, type of flow according to the relative chronological position of the source and target episode

[9] We summarize and group nationalities according to differences in their respective legal situation into four nationality-groups „Austria", „EU-Neu", „EU-Alt", „Drittstaat"

[10] EU-Neu comprises: Czech Republic, Hungary, Poland, Slovakia,...

Figure 2. Chord diagram interactive visualization 2011, source: IHS calculations based on official insurance data

4. DISCUSSION

When we started working with the data in 2008 it was immediately evident that it would give us a lot more discretion in extracting and preparing the data for traditional methods and models of labor market analysis such as discrete choice models for modelling retirement decisions or methods for program evaluation.

In our work adapting Information Technology Processes such as KDD [9] or more recent Data Science [10], has helped us develop new insights as shown in this work.[11]

As our two examples demonstrate: visualizations, simple ones as well as more complex and interactive ones, play an important role in our workflow: for inspecting the data, confirming or rejecting hypothesis and developing new ones, but also for disseminating and communicating results. Appropriate visualizations have a tremendous impact for getting a message across to a non-technical or non-scientific audience. For instance, in case of the lay-off tax we could communicate vital aspects to the legislators.

It is important to point out however, that the data we use represent an electronic record of the non-technically structured administrative processes of Austrian social insurance system. The time and effort necessary to understand, extract, preprocess and transform the raw data is considerable. In case of the interactive visualization we used about 60 percent of our time for the data transformation process. This is in line with ranges of 50 to 80 percent published in newspapers[12] and under the 80 percent reported in Kandel [11].

5. FUTURE WORK

Currently we are finishing the project of Stock-Flow-Consistent data framework for analysis of employment. One of the remaining issues is the identification of flows that are artefacts of

[11] From the technology side we needed: PostgreSQL server for data handling, a mix of SQL and Python code for table scripting, d3 & java script & html for interactive chord diagrams and a webserver for making it accessible over the internet.

[12] http://mobile.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=3

administrative idiosyncrasies of the Austrian Social Insurance System.

We will continue implementing methods from the realm of Data Science including visualizations as a complementary approach to traditional economic models.

Currently we are evaluating methods from graph/network theory in order to see the Austrian social insurance data as a graph and predict future employment spells and flows [12].

6. REFERENCES

[1] Hanappi, T. and Müllbacher, S. and Heil, E., Auswirkungen von finanziellen Anreizen auf das Pensionsantrittsalter. Eine Analyse mit dem Mikrosimulationsmodell IREA ,Studie im Auftrag des Bundesministeriums für Finanzen, Research Report, 2014

[2] Walch, D. and Segert, A. and Heil, E., Erwerbstätige MigrantInnen im Tourismus, Studie im Auftrag des Bundesministerium für Wissenschaft, Forschung und Wirtschaft, Research Report, 2012

[3] Schwarzbauer, W. and Hanappi, T. and Heil, E. and Hofer, H. and Hyee, R. and Müllbacher, S. and Schuster, P., Analyse einer Erhöhung des effektiven Pensionsantrittsalters, Studie im Auftrag des Bundesministeriums für Finanzen. Research Report, 2012

[4] Davis,S. J. and Faberman R. J. and Haltiwanger, J., The Flow Approach to Labor Markets: New Data Sources and Micro–Macro Links, *Journal of Economic Perspectives. 20:3*, Pages 3–26, Summer 2006

[5] Burgess, S. and Lane J. I. and Stevens, D., Job flows, Worker Flows, and Churning, *Journal of Labor Economics. 18:3*, pp. 473–502. 2000

[6] Bjelland, M. and Fallick, B. and Haltiwanger, J. and McEntarfer, E., Employer-to-Employer Flows in the United States: Estimates Using Linked Employer-Employee Data, *Journal of Business & Economic Statistics, 29:4*, 493-505, 2011

[7] Heil E. and Walch W., A Flow Consistent Data Framework for the Analysis of Employment using Administrative Data from the social insurance system, mimeo

[8] Abel, G.J. and Sander, N., Quantifying global international migration flows,*Science*,*343*,6178,1520-1522, American Association for the Advancement of Science,2014

[9] Fayyad, U. and Piatetsky-Shapiro, G. and Smyth, P., The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM 39, 11*, 27-34. DOI=10.1145/240455.240464, November 1996 http://doi.acm.org/10.1145/240455.240464

[10] Dhar, V., Data science and prediction, *Communications of the ACM 56 (12): 64*, 2013, http://doi.acm.org/10.1145/2500499

[11] Kandel, S. and Heer, J. and Plaisant, C. and Kennedy, J. and van Ham, F. and Riche, N.H. and Weaver, C. and Lee, B. and Brodbeck, D. and Buono, P., Research directions in data wrangling: Visualizations and transformations for usable and credible data, *Information Visualization Journal, 10(4)*, 271–288, 2011

[12] Liben-Nowell, D. and Kleinberg, J. (2007), The link-prediction problem for social networks, *J. Am. Soc. Inf. Sci., 58*: 1019–1031. doi: 10.1002/asi.20591

.