Innovation Action (IA)

ICT-14-2016-2017

H2020-ICT-2017-1

Enabling procurement data value chains for economic development, demand management, competitive markets and vendor intelligence

THEY BUY FOR YOU

Deliverable D3.3

D3.3 Public spending real-time monitoring and analytics demo v1

Date	21/03/2019
Author(s)	Matej Kovačič, Matej Posinković (JSI), Carlos Badenes-Olmedo (UPM)
Dissemination level	Public
Work package	3
Version	5

Document metadata

Quality assurers and contributors

Quality assurer(s)	Ahmet Soylu, Till Christopher Lech (SINTEF)
Contributor(s)	Matej Kovačič, Matej Posinković, Carlos Badenes-Olmedo

Version history

Date	Version	Description
18/03/2019	1	Initial report submitted for review
21/03/2019	2	Added code examples and some screenshots
9/4/2019	3	Added document comparison and more detailed description of web platform. Added chapter on document similarity.
11/4/2019	4	Added a more detailed description of the search algorithm for multilingual documents.
12/4/2019	5	Checked spelling and grammar.
12/4/2019	Final	Final version for submission

Executive summary

This report describes deliverable D3.3, with goal to develop an initial implementation of the public spending monitoring framework implemented as web application, including clustering of related documents across languages and visualisations of cluster data.

Table of contents

1 Objectives

The main goal of work package WP3 is to develop methodology and tools for automatic real-time monitoring and analysis of public spending data.

D3.3 is highly intertwined with D3.2, since the application code developed in D3.2 is designed as a web - oriented application from the start.

2 Approach

Our approach towards development of public spending monitoring framework is divided into three areas. The first part of the web application will be used to do very basic analysis and visualisation of public spending data and public procurement data.

The second part is directed towards anomaly detection and visualisation of detected anomalies in spending data. We have implemented five methods of anomaly detection in data as a web application, namely Average Deviation Anomaly, Jenks Natural Breaks, Period Margin Points Cumulatives, Local Extremes Detection and Time Periods Deviations. These methods were developed in D3.2.

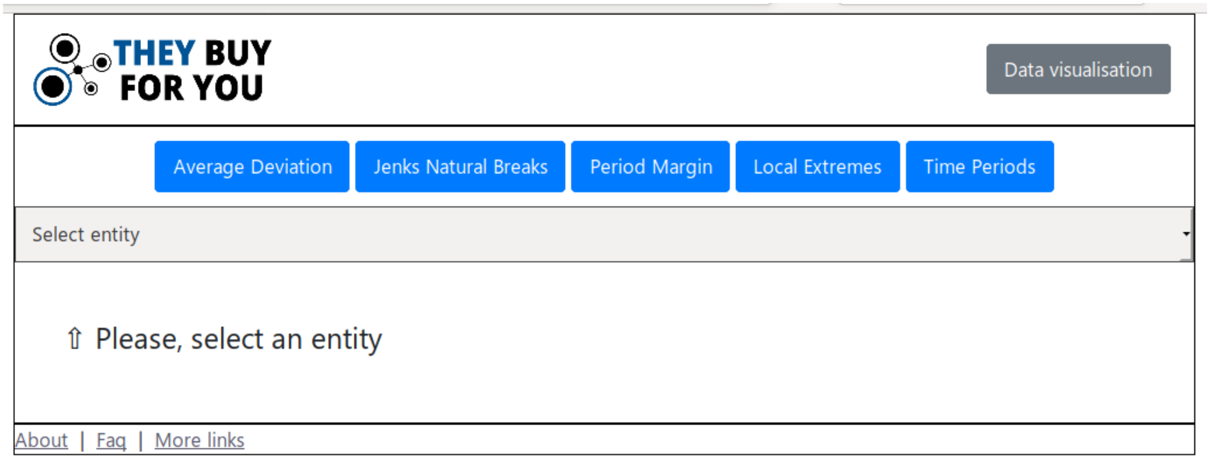
The third part is related towards clustering of documents across languages, i. e. document similarity. For document similarity analysis we have used two approaches, which will be described later on. This part has been developed partly in D3.1 and partly is developed by our partners from Universidad Politécnica de Madrid.

3 Application

We have developed initial implementation of the public spending monitoring framework as a web application.

Analysis and visualisation of data

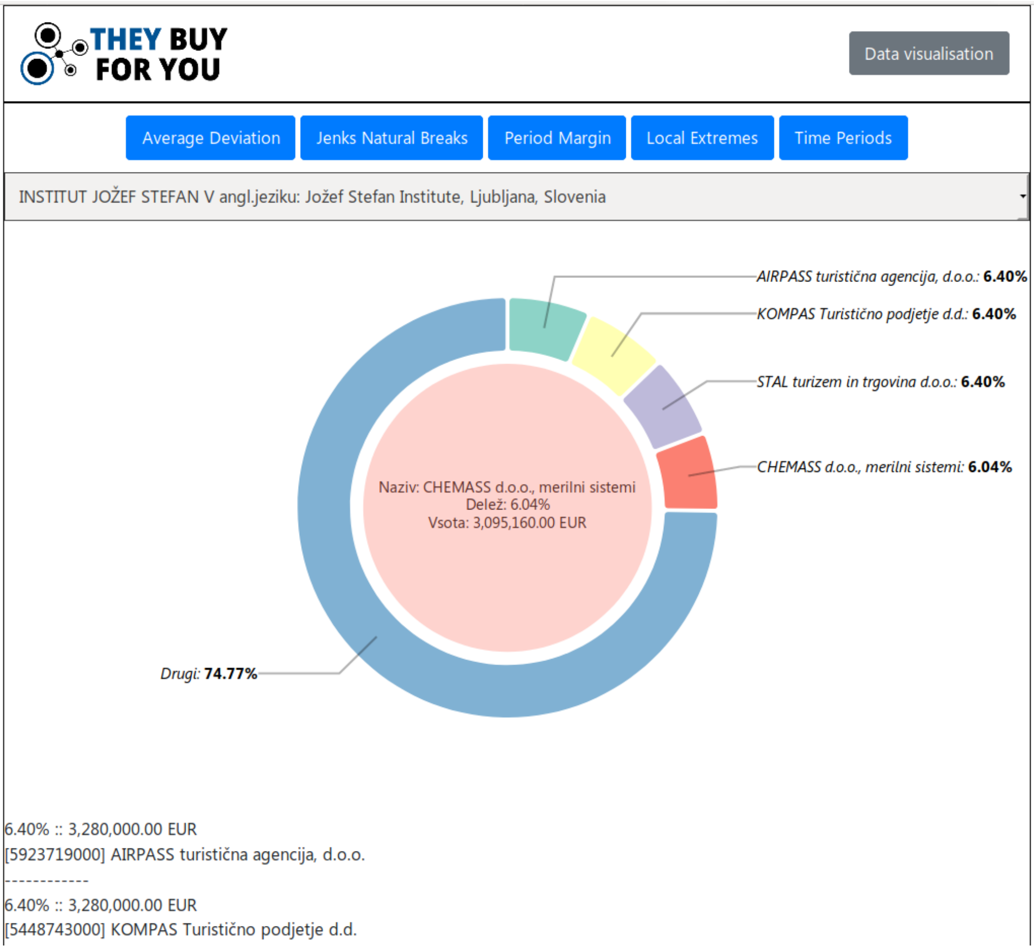
The first part of an application is being used for analysis and visualisation of public spending data and public procurement data and also for displaying groups of entities with detected anomalies and visualisation of detected anomalies. We have implemented basic search service among the data, however this part will be further improved.



⇧ Please, select an entity

If the user selects "Data visualisation" he or she can then select public sector entity and get an overview of transactions from this public sector entity to private companies. Currently data are just printed out and some very basic visualisation has been implemented. The user can also search among institutions by typing the name of the institution.

Example of financial transactions (payments) of Jožef Stefan Institute to private companies:


THEY BUY FOR YOU
Data visualisation
Average Deviation
Jenks Natural Breaks
Period Margin
Local Extremes
Time Periods
AIRPASS turistična agencija, d.o.o.: 6.40%
KOMPAS Turistično podjetje d.d.: 6.40%
STAL turizem in trgovina d.o.o.: 6.40%
CHEMASS d.o.o., merilni sistemi: 6.04%
Naziv: CHEMASS d.o.o., merilni sistemi
Delež: 6.04%
Vsota: 3,095,160.00 EUR
Drugi: 74.77%

INSTITUT JOŽEF STEFAN V angl.jeziku: Jožef Stefan Institute, Ljubljana, Slovenia

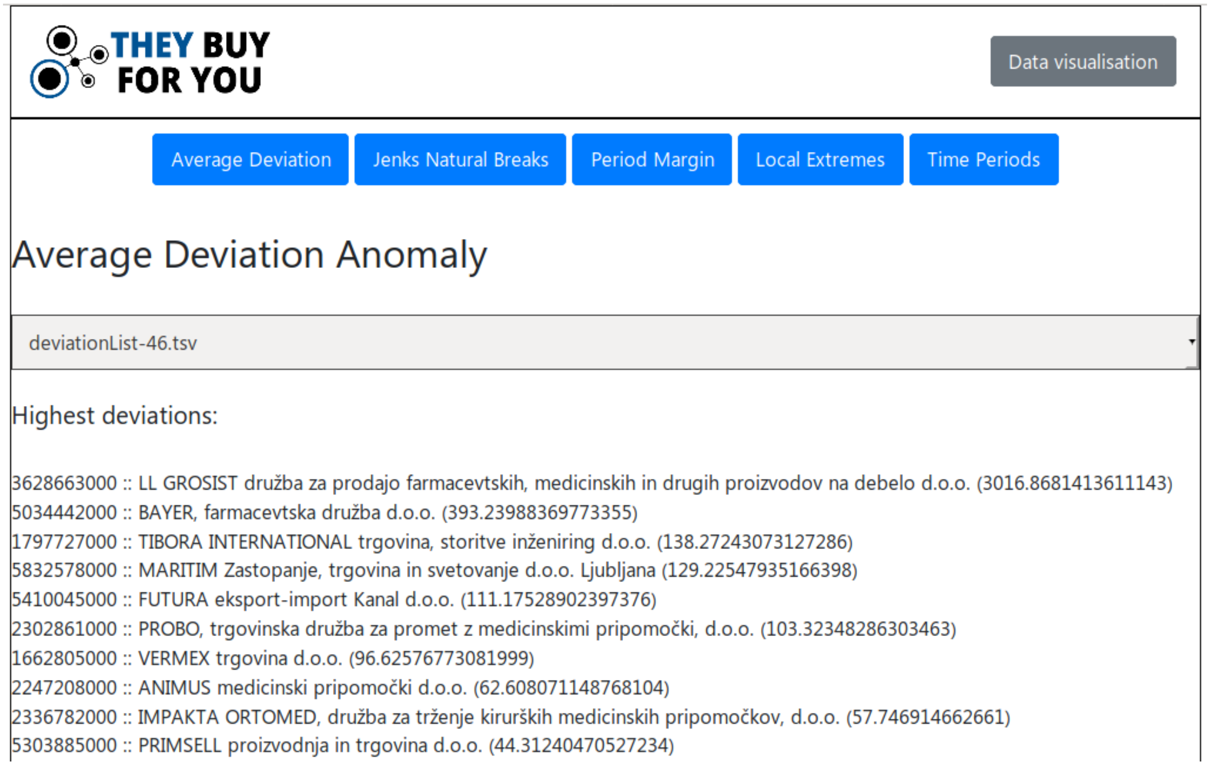
```
6.40% :: 3,280,000.00 EUR
[5923719000] AIRPASS turistična agencija, d.o.o.
------------
6.40% :: 3,280,000.00 EUR
[5448743000] KOMPAS Turistično podjetje d.d.
```

Anomaly detection and visualisation of detected anomalies

The second part of the application is implementation of anomaly detection and visualisation of detected anomalies in spending data.

We have implemented five methods of anomaly detection: Average Deviation Anomaly, Jenks Natural Breaks, Period Margin Points Cumulative, Local Extremes Detection and Time Periods Deviations. The user can select different anomaly detection methods by clicking to respective buttons. Anomalies are displayed in the whole data or in different industry sectors only. The latter means, that anomaly detection method is using only anomalies in transactions among companies registered for the specific activity (building, IT, agriculture, …).

There is an example of a printout of the highest deviations for companies with main activity selling of pharmaceutical products with Average Deviation anomaly detection method:



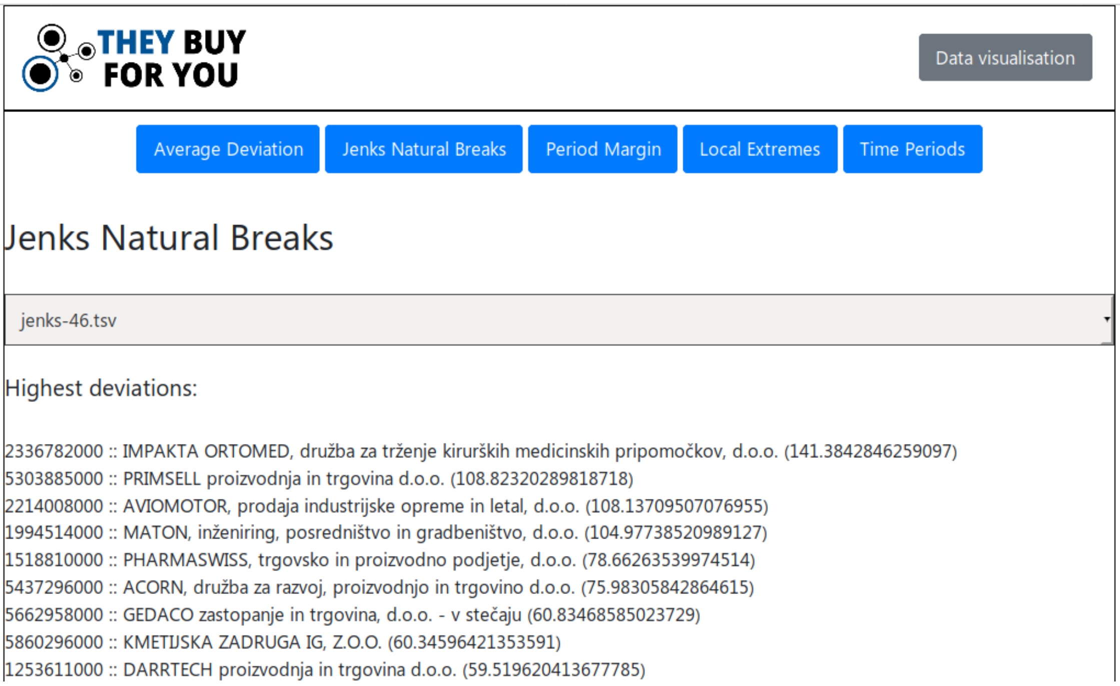
Average Deviation Anomaly

deviationList-46.tsv

Highest deviations:

3628663000 :: LL GROSIST družba za prodajo farmacevtskih, medicinskih in drugih proizvodov na debelo d.o.o. (3016.8681413611143)
5034442000 :: BAYER, farmacevtska družba d.o.o. (393.23988369773355)
1797727000 :: TIBORA INTERNATIONAL trgovina, storitve inženiring d.o.o. (138.27243073127286)
5832578000 :: MARITIM Zastopanje, trgovina in svetovanje d.o.o. Ljubljana (129.22547935166398)
5410045000 :: FUTURA eksport-import Kanal d.o.o. (111.17528902397376)
2302861000 :: PROBO, trgovinska družba za promet z medicinskimi pripomočki, d.o.o. (103.32348286303463)
1662805000 :: VERMEX trgovina d.o.o. (96.62576773081999)
2247208000 :: ANIMUS medicinski pripomočki d.o.o. (62.608071148768104)
2336782000 :: IMPAKTA ORTOMED, družba za trženje kirurških medicinskih pripomočkov, d.o.o. (57.746914662661)
5303885000 :: PRIMSELL proizvodnja in trgovina d.o.o. (44.31240470527234)

An example of a printout of deviations for companies with main activity selling of pharmaceutical products with Jenks Natural Breaks anomaly detection method:



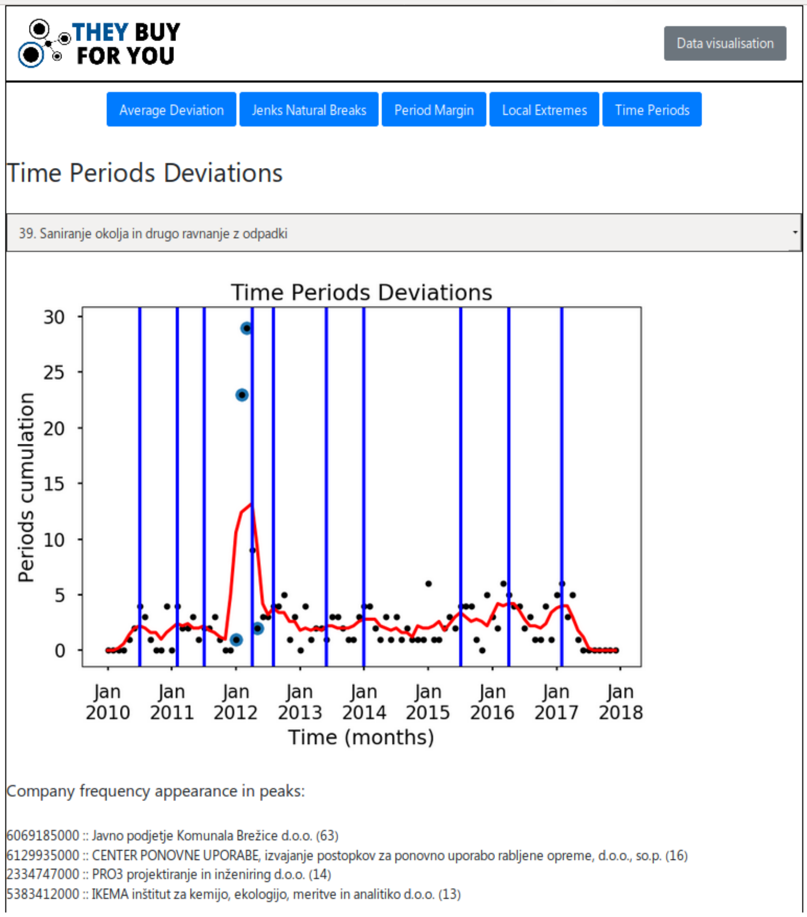
Jenks Natural Breaks

jenks-46.tsv

Highest deviations:

2336782000 :: IMPAKTA ORTOMED, družba za trženje kirurških medicinskih pripomočkov, d.o.o. (141.3842846259097)
5303885000 :: PRIMSELL proizvodnja in trgovina d.o.o. (108.82320289818718)
2214008000 :: AVIOMOTOR, prodaja industrijske opreme in letal, d.o.o. (108.13709507076955)
1994514000 :: MATON, inženiring, posredništvo in gradbeništvo, d.o.o. (104.97738520989127)
1518810000 :: PHARMASWISS, trgovsko in proizvodno podjetje, d.o.o. (78.66263539974514)
5437296000 :: ACORN, družba za razvoj, proizvodnjo in trgovino d.o.o. (75.98305842864615)
5662958000 :: GEDACO zastopanje in trgovina, d.o.o. - v stečaju (60.83468585023729)
5860296000 :: KMETIJSKA ZADRUGA IG, Z.O.O. (60.34596421353591)
1253611000 :: DARRTECH proizvodnja in trgovina d.o.o. (59.519620413677785)

An example of visualisation of anomalies in transactions to companies registered for waste management with Time Periods Deviations anomaly detection method:



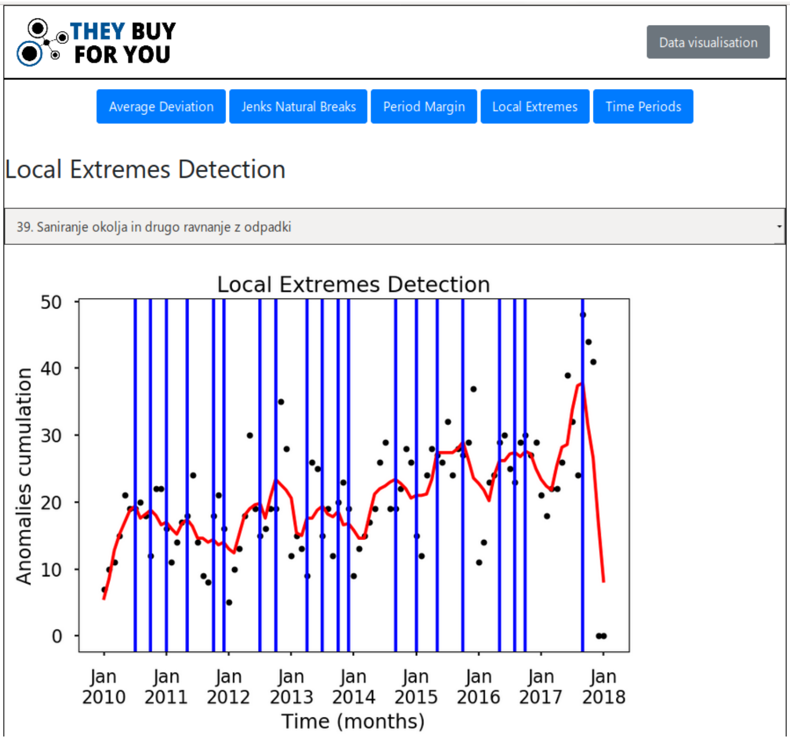
Time Periods Deviations

39. Saniranje okolja in drugo ravnanje z odpadki

Company frequency appearance in peaks:

6069185000 :: Javno podjetje Komunala Brežice d.o.o. (63)
6129935000 :: CENTER PONOVNE UPORABE, izvajanje postopkov za ponovno uporabo rabljene opreme, d.o.o., so.p. (16)
2334747000 :: PRO3 projektiranje in inženiring d.o.o. (14)
5383412000 :: IKEMA inštitut za kemijo, ekologijo, meritve in analitiko d.o.o. (13)

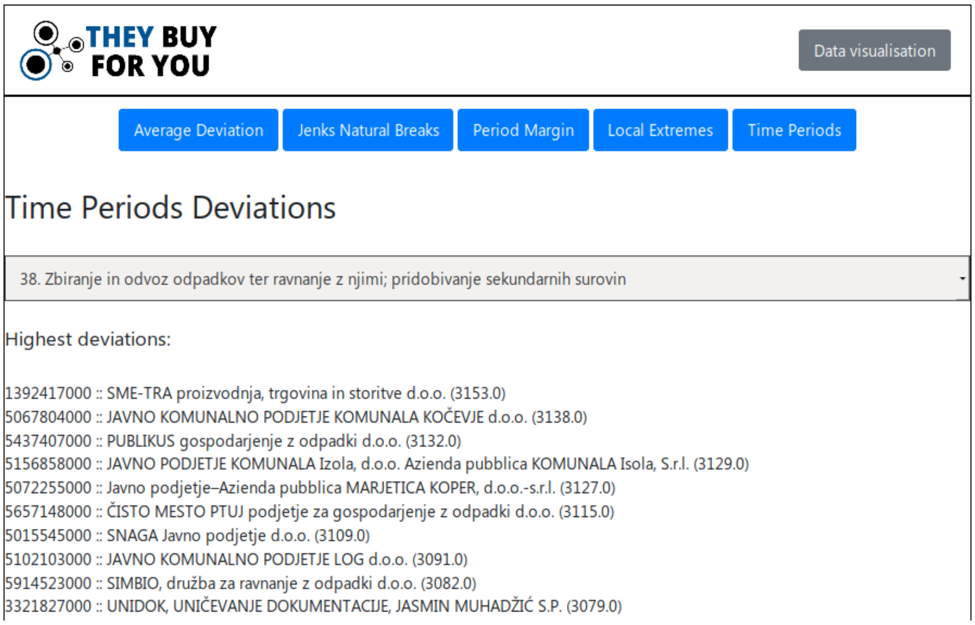
An example of visualisation of anomalies in transactions to companies registered for waste management with Local Extremes Detection anomaly detection method:



Local Extremes Detection

39. Saniranje okolja in drugo ravnanje z odpadki

An example of visualisation of anomalies in transactions to companies registered for waste management with Time Period Deviations anomaly detection method:



Time Periods Deviations

38. Zbiranje in odvoz odpadkov ter ravnanje z njimi; pridobivanje sekundarnih surovin

Highest deviations:

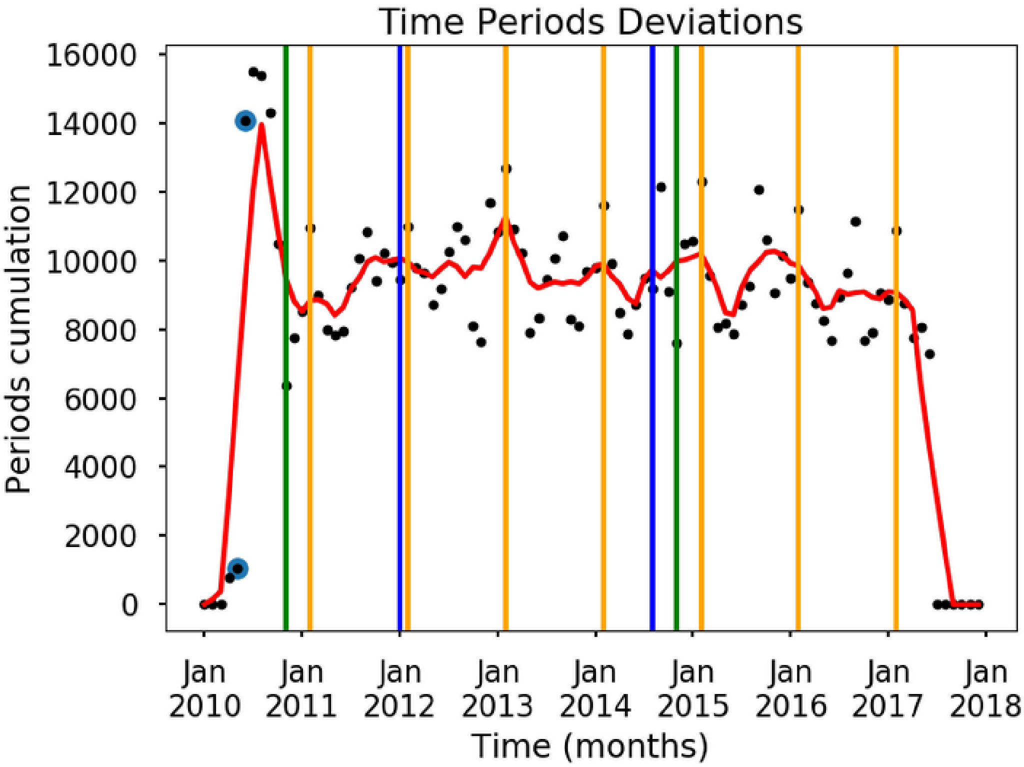
1392417000 :: SME-TRA proizvodnja, trgovina in storitve d.o.o. (3153.0)
5067804000 :: JAVNO KOMUNALNO PODJETJE KOMUNALA KOČEVJE d.o.o. (3138.0)
5437407000 :: PUBLIKUS gospodarjenje z odpadki d.o.o. (3132.0)
5156858000 :: JAVNO PODJETJE KOMUNALA Izola, d.o.o. Azienda pubblica KOMUNALA Isola, S.r.l. (3129.0)
5072255000 :: Javno podjetje–Azienda pubblica MARJETICA KOPER, d.o.o.-s.r.l. (3127.0)
5657148000 :: ČISTO MESTO PTUJ podjetje za gospodarjenje z odpadki d.o.o. (3115.0)
5015545000 :: SNAGA Javno podjetje d.o.o. (3109.0)
5102103000 :: JAVNO KOMUNALNO PODJETJE LOG d.o.o. (3091.0)
5914523000 :: SIMBIO, družba za ravnanje z odpadki d.o.o. (3082.0)
3321827000 :: UNIDOK, UNIČEVANJE DOKUMENTACIJE, JASMIN MUHADŽIĆ S.P. (3079.0)

4 Anomaly detection interpretation

The purpose of anomaly detection is usually the identification of rare events or unexpected bursts in activity, which raise suspicions by differing significantly from the majority of the data. In the area of public spending, there is a high probability, that anomalous items will translate to some kind of problem such as structural defect, illicit management, abuse of functions, unfair competitiveness, and clientelism or some form of direct corruption.

For assessment of the efficiency of our anomaly detection methods, we would need the support of domain experts, however, we have already found some interesting patterns in public spending, which we present below.

The first example is the analysis of all Slovenian public spending data with Time Periods Deviations method, which has detected some interesting patterns in public spending changes.



The x-axes represent time. It starts in January in 2010, where one unit means one month. Y-axis represents a number of detected "cases" in this case, start/end of the transaction period.

The extremes clearly coincide with dates:

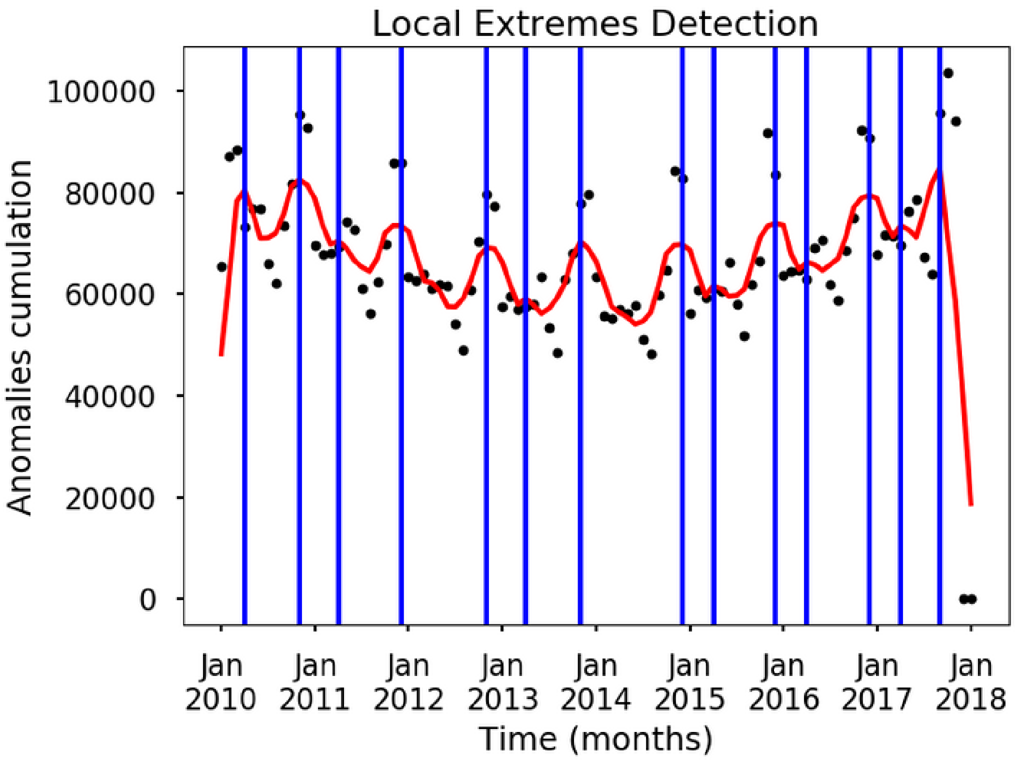
- orange lines: every January (and the beginning of the budget year),
- blue line: local elections,
- green line: parliamentary elections.

Next to that there are some other extremes visible:

- around the first blue line Slovenia was experiencing a political and economic crisis; that was the period when governments were rapidly changing,
- extremes around 2016, where Slovenia was experiencing a migrant crisis.

The second example is the analysis of all Slovenian public spending data with Local Extremes Detection method, which has also detected some interesting patterns in public spending changes.

Copyright© TheyBuyForYou Consortium 2018-2020



The x-axes represent time. It starts in January in 2010, where one unit means one month. Y-axis represents a number of detected "cases" in this case, the number of anomalous transaction extremes.

There are three interesting observations:

- The graph has a global minimum around the summer of 2014, when Slovenia finally overcame the economic crisis. This can be explained as a consequence of government savings programmes.
- Extremes on the second graph roughly coincide to the first graph, which is expected as methods share some approaches.
- Specific extremes are associated to certain events: October 2010 (big floods in Slovenia in the previous month), November 2011 (national elections), October 2012 (big floods), November 2014 (local elections).

  We have not (yet) tied external events to all identified extremes. However, we have noticed a large jump pattern, followed by a small jump.

5 Used data

Our application is currently using only Slovenian data for testing. We plan to do some tests with other partner's data, but the final version will enable other partners to upload their data into our framework and our application will be able to do anomaly detection analysis on their data sets.

Currently we are using the following data:

Slovenian Business Register, including historical records

Slovenian Business Register is maintained by AJPES ("Agency of the Republic of Slovenia for Public Legal Records and Related Services."). It contains all companies (public or private) in Slovenia alongside their owners and roles within companies.

Next to the current database of all companies in Slovenia, we have all historical entries (i.e., companies that have already been shut down), which help us track relations that existed in the past.

Description of the data is available on this URL:
https://www.ajpes.si/Doc/AJPES/Za_razvijalce/Struktura_podatkov_PRS_042016.doc

Data structure is described in this document:
https://www.ajpes.si/Doc/Registri/PRS/The_table_structure_of_the_Slovenian_Business_Register.doc

The data are distributed in MDB format (MS Access). The application for importing the data into PostgreSQL is available on Github: https://github.com/zejn/mdb2csv.

The database is available via Access to Public Information Act, which has been a very time-consuming procedure. Ajpes agency is known to be very unresponsive to access to public information requests, since they are also selling the data from Slovenian Business Register. They refused our request several times and after several complaints we finally managed to get the data.

Slovenian spending data from Erar.si

Erar database consists of all financial transactions between public entities and privately held companies (i. e. spending data). The data are publicly available on a monthly basis.

Slovenian spending data are available to download at the following URL:
http://erar.si/cdn/podatki/

Description of the data is available here:
https://sintef.sharepoint.com/teams/work-2368/Shared%20Documents/Work/Deliverables/D3.2%20Common%20spending%20templates%20framework%20v1/Description_of_Erar_dataset.pdf

The database is available for download for free and without registration. Data are updated monthly and could be automatically gathered and imported.

For gathering the data we are using simple bash script:

```
for i in {2003..2017}; do wget http://erar.si/cdn/podatki/trans$i.csv.gz;
done

for i in {2003..2017}; do gunzip trans$i.csv.gz; done
```

The data from the last year (i. e. 2018), are on Erar available on a monthly basis, so the example of gathering script is:

```
for i in {01..12}; do wget http://erar.si/cdn/podatki/trans2018$i.csv.gz;
done

for i in {01..12}; do gunzip trans2018$i.csv.gz; done
```

The data are then imported into the PostgreSQL database. From there we import the data into our internal structure used for anomaly detection. The application code is available on Github:

https://github.com/TBFY/transactions-data-translator

Standard Classification of Activities database

Standard Classification of Activities database is the obligatory national standard used for defining the main activity and for classifying business entities and their units for the needs of official and other administrative data collections (registers, records, databases, etc.) and for the needs of national and international statistics and analyses. In line with Article 6 of the Decree on the 2008 Standard Classification of Activities, the Statistical Office of the Republic of Slovenia is authorised to explain the content of classification items. Classification of units of the Business Register of Slovenia by activity is the responsibility of the Agency of the Republic of Slovenia for Public Legal Records and Related Services (AJPES).

Standard Classification of Activities database is available from this URL:
https://www.stat.si/Klasje/Klasje/Tabela/5531

The database does not change, so there is no need for updating.

Public procurement data from the Ministry of Public Administration

The Ministry of Public Administration has provided us a complete database of public procurement for years 2015, 2016 and 2017. The database consists of publication of all public tenders (public procurements) and information about signed contracts and framework agreements.

The public procurement database description of fields is uploaded to the project’s Sharepoint:
https://sintef.sharepoint.com/:x:/r/teams/work-2368/_layouts/15/Doc.aspx?sourcedoc=%7B3FD6703F-8321-4110-A6A1-0A3391653628%7D&file=Procurement_MJU_column_names.xlsx&action=default&mobileredirect=true

We have requested Public procurement data from the Ministry of Public Administration and received them very quickly. The Ministry of Public Administration will also develop the API to share the data in a much easier and quick way with the interested public.

Registry of budget users

Registry of budget users is available in Excel format from this URL:
https://www.ujp.gov.si/dokumenti/dokument.asp?id=316

The database does change very rarely (when a new public entity is established or closed, so the data are manually updated and imported into the database.

Partner's data

We have obtained access to OpenOpps spending data. In the following months, the data will be imported into our internal database and we will test our anomaly detection methods with this data.

6 Application code

The application code is published on our Github repository:
https://github.com/TBFY/web-application

The application itself is currently available on the following URL: http://tbfy.ijs.si/

7 Future work in the area of anomaly detection

Our future work will be oriented towards further development of visualisation and search among the data. In the final application we will also need to improve UX design.

We are also developing API, which will allow other partners to analyse their data with our application and to include generated visualisations into their summaries and infographics.

We are planning to add some additional specific analyses for public procurement data.

8 Clustering of documents across languages

For document similarity analysis we have developed two approaches. One is to compute similarity between two given documents and the other is development of search algorithm for multilingual documents. These approaches are not part of the web framework yet, however, similarity computation service API is already available and running, a description of which is provided in Deliverable 5.3: Procurement APIs and Platform Release v2.

Similarity between documents

Since input is text documents in different languages, we must compute the similarity with removed effect of the language. There are several methods to compute similarity from multilingual documents: statistical cross-lingual computation, semantic cross-lingual computation and similarity computation through machine translation. We have implemented all three methods in our document similarity computation service in D3.1.

Our approach is to use (1) Wikipedia concepts (for input document we obtain a set of Wikipedia concepts relevant to that document and these concepts are mapped into English-language equivalents using the Wikipedia's cross-language links; finally we compute various measures of similarity between the resulting sets of annotations for each input document; (2) CCA projections (documents in different languages are represented by vectors in different high-dimensional spaces and cannot be compared directly; with CCA (canonical correlation analysis) we compute projections from these spaces into a new, shared, language-independent space) and (3) translations (we use machine-translation service to translate both input documents into English).

This technique is useful for comparing two documents but turned out not to be very effective when we want to find all similar documents if number of documents is quite large.

The search algorithm for multilingual documents

Most textual search engines are based on patterns. They identify the keywords in the query that will be used to filter and sort the results by relevance. This prevents texts written in languages other than the query itself from being retrieved without prior translation. To solve this problem, recent algorithms define a single representational space shared by all languages where the searches are carried out. These approaches, however, are not suitable for working with large collections of documents as they do not use reduced representational spaces.

This approach addresses the problem of large-scale multilingual searches by proposing an algorithm based on hierarchical annotations from probabilistic topic models aligned between multiple languages. Documents, regardless of their language, are then described by hash expressions that allow to retrieve similar documents and to perform thematic searches without the need for any translation. The hashing algorithm used for large-scale document search considers relative hierarchical thresholds for each relevance level. The topic distributions are then described by points in a single dimension where topics closely packed together are grouped in the feature space. This approach does not require a fixed number of groups. It only requires a maximum distance (eps) to consider two points close and grouped, that can be estimated from the own distribution of topics (e.g., variance). More details about the algorithm are described in the paper "*Efficient Exploration of Scientific Articles using Topic-based Hashing Algorithms,*" which is currently under review in the Semantic Web Journal[1].

[1] http://www.semantic-web-journal.net/content/efficient-exploration-scientific-articles-using-topic-based-hashing-algorithms

http://www.semantic-web-journal.net/content/efficient-exploration-scientific-articles-using-topic-based-hashing-algorithms

The multilingual capability of the algorithm is obtained by aligning probabilistic topic models created from multi-lingual documents annotated with shared labels. This results in a single representational space shared by all languages where documents are described by hash expressions of hierarchical topics.

Taking into account these annotations, searches can be done from a given document or even from free text, in any language. Extensive evaluations carried out with the parallel corpus JRC-Acquis[2] and the multilingual thesaurus Eurovoc[3] validate the proposed algorithm, offering very promising results with precision values close to 0.9.

For now, the multilingual capability of the algorithm could be described as an evolution of the previous algorithm using aligned probabilistic topic models from multi-lingual documents annotated with shared labels. This results in a single representational space shared by all languages where documents are described by hash expressions of hierarchical topics.

[2] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[3] https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc