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 D3.2

D3.2 Common spending templates framework v3

Date	21/03/2019
Author(s)	Matej Kovačič, Matej Posinković (JSI)
Dissemination level	Public
Work package	3
Version	4

Document metadata

Quality assurers and contributors

Quality assurer(s)	Ahmet Soylu, Till Christopher Lech (SINTEF)
Contributor(s)	Matej Kovačič, Matej Posinković (JSI)

Version history

Date	Version	Description
18/03/2019	1	Initial report submitted for review
21/03/2019	2	Added code samples published on Github, naming of the methods, added chapter on data importing, other small changes.
8/04/2019	3	Added more detailed description of methods and description of data.
12/04/2019	4	Spell and grammar checking.
12/04/2019	Final	Final version for submission

Executive summary

This report describes deliverable D3.2 with goal to develop an initial version of the framework for anomaly detection in public spending, including the discovery of typical groups or sequences of public spending orders.

It provides a brief description of technology for anomaly detection that has been implemented and the results of the initial testing on Slovenian public spending data.

Table of contents

1 Objectives

The main goal of work package 3 (WP3) is to develop methodology and tools for automatic real-time monitoring and analysis of public spending data.

The key contribution to real-time monitoring of spending data is a software system for analytical anomaly detection from structured data on transactions in public spending data streams. The data consists of (a) a stream of financial transactions between entities, and (b) additional meta data on entities themselves.

In the first phase, financial transactions have been converted into a dynamic network encoding a flow of money across the entities in the ecosystem. Such data structure serves as a basis for follow-up analyses. The aim is to transform transaction data and corresponding meta data on entities into a data structure which allows complex analysis and per-request query operators. The query operators allow detecting unusual more or less complex patterns in spending patterns across the network and alarming in business real-time on the detected suspicions situations.

The precise technology for anomaly detection has been implemented as a transformation of the dynamic network into sparse feature vectors, which has been further used with more traditional machine learning techniques (either supervised or unsupervised) to detect anomalous situations. The key to detecting relevant signals in the data is in the representation of the transformed vectors where we combined human experience (in the form of which features to extract) with statistical/ML techniques (detecting unusual behaviour in the data).

The system is highly scalable (i.e. to process in business real-time tens of millions of transactions) and allows detecting a large class of anomalies in the automatic mode or in the exploratory mode (with human-machine interaction).

2 Methods of anomaly detection

We have developed and implemented several anomaly detection methods:

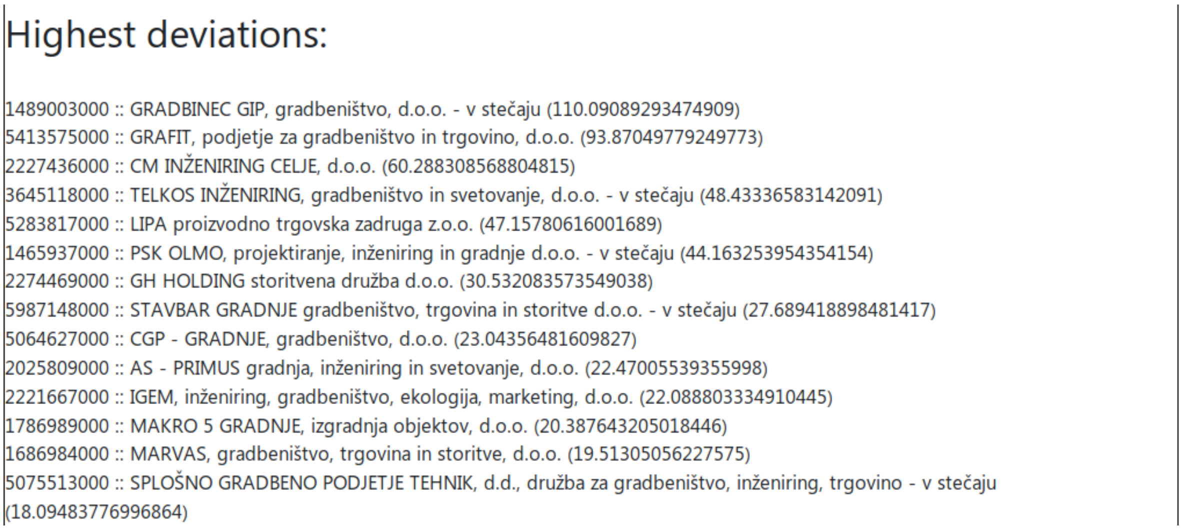
Method 1: Average Deviation Anomaly

The first method summarises financial transactions between two entities and within the pool of all two-entity transaction sum, find the most deviating ones. We have applied this method on all entities as well as on entities grouped based on company classifiers (e.g. separate construction oriented companies from IT-oriented companies). For classification of companies, we are using the Standard The classification of Activities, which is the obligatory national standard in Slovenia and is used for defining the main activity of business entities and their units.

Example of method's core Python code:

```
def createAverageDeviationList(self, adConfig):

# get params
# get params

file2AnalysePath = adConfig['files2AnalysePath'] if 'files2AnalysePath' in adConfig else
''
files2AnalyseName = adConfig['files2AnalyseName'] if 'files2AnalyseName' in adConfig else
''
tmp_filePath = file2AnalysePath + files2AnalyseName

# get sorted average deviations
# get sorted average deviations

self.readTransactionsFileData(tmp_filePath)
self.deviationList = self.getSortedDeviationList(self.transactionsDataSums)
for classifier in self.transactionsDataSumsClassified:
self.deviationListClassified[classifier]                                           =
self.getSortedDeviationList(self.transactionsDataSumsClassified[classifier])

return 0
```

Example of detected deviations in transactions from budget users to private companies registered to build construction:



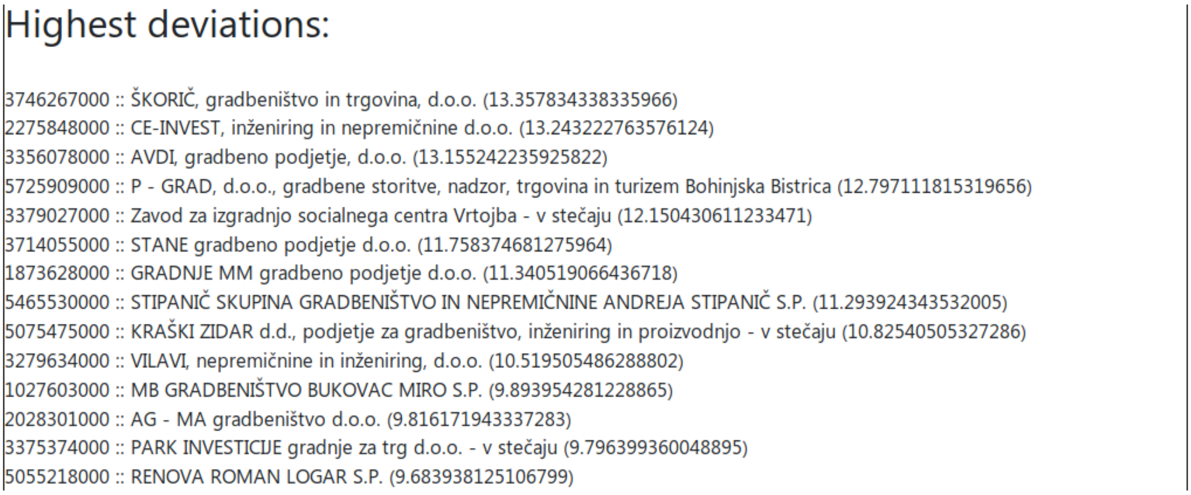
```
Highest deviations:

1489003000 :: GRADBINEC GIP, gradbeništvo, d.o.o. - v stečaju (110.09089293474909)
5413575000 :: GRAFIT, podjetje za gradbeništvo in trgovino, d.o.o. (93.87049779249773)
2227436000 :: CM INŽENIRING CELJE, d.o.o. (60.288308568804815)
3645118000 :: TELKOS INŽENIRING, gradbeništvo in svetovanje, d.o.o. - v stečaju (48.43336583142091)
5283817000 :: LIPA proizvodno trgovska zadruga z.o.o. (47.15780616001689)
1465937000 :: PSK OLMO, projektiranje, inženiring in gradnje d.o.o. - v stečaju (44.163253954354154)
2274469000 :: GH HOLDING storitvena družba d.o.o. (30.532083573549038)
5987148000 :: STAVBAR GRADNJE gradbeništvo, trgovina in storitve d.o.o. - v stečaju (27.689418898481417)
5064627000 :: CGP - GRADNJE, gradbeništvo, d.o.o. (23.04356481609827)
2025809000 :: AS - PRIMUS gradnja, inženiring in svetovanje, d.o.o. (22.47005539355998)
2221667000 :: IGEM, inženiring, gradbeništvo, ekologija, marketing, d.o.o. (22.088803334910445)
1786989000 :: MAKRO 5 GRADNJE, izgradnja objektov, d.o.o. (20.387643205018446)
1686984000 :: MARVAS, gradbeništvo, trgovina in storitve, d.o.o. (19.51305056227575)
5075513000 :: SPLOŠNO GRADBENO PODJETJE TEHNIK, d.d., družba za gradbeništvo, inženiring, trgovino - v stečaju
(18.09483776996864)
```

Method 2: Jenks Natural Breaks

The second method's approach is the same as Method 1, except, it organizes transaction sums into an optimal number of clusters and define deviations within each cluster separately.

This method is a data clustering method designed to determine the best arrangement of values into different classes. The method is seeking to minimize each class's average deviation from the class mean and at the same time it is maximizing each class's deviation from the means of the other groups. So method cluster data in a manner that it reduces the variance within classes and maximize the variance between classes.

Example of detected deviations in transactions from budget users to private companies registered to build construction:

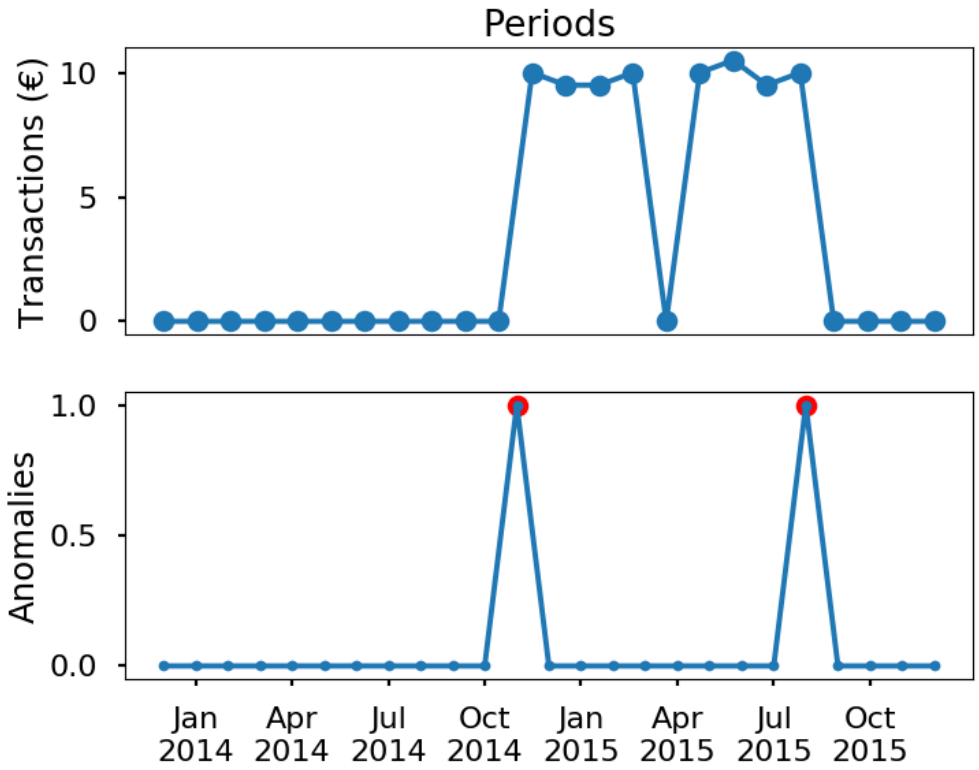


```
Highest deviations:

3746267000 :: ŠKORIČ, gradbeništvo in trgovina, d.o.o. (13.357834338335966)
2275848000 :: CE-INVEST, inženiring in nepremičnine d.o.o. (13.243222763576124)
3356078000 :: AVDI, gradbeno podjetje, d.o.o. (13.155242235925822)
5725909000 :: P - GRAD, d.o.o., gradbene storitve, nadzor, trgovina in turizem Bohinjska Bistrica (12.797111815319656)
3379027000 :: Zavod za izgradnjo socialnega centra Vrtojba - v stečaju (12.150430611233471)
3714055000 :: STANE gradbeno podjetje d.o.o. (11.758374681275964)
1873628000 :: GRADNJE MM gradbeno podjetje d.o.o. (11.340519066436718)
5465530000 :: STIPANIČ SKUPINA GRADBENIŠTVO IN NEPREMIČNINE ANDREJA STIPANIČ S.P. (11.293924343532005)
5075475000 :: KRAŠKI ZIDAR d.d., podjetje za gradbeništvo, inženiring in proizvodnjo - v stečaju (10.82540505327286)
3279634000 :: VILAVI, nepremičnine in inženiring, d.o.o. (10.519505486288802)
1027603000 :: MB GRADBENIŠTVO BUKOVAC MIRO S.P. (9.893954281228865)
2028301000 :: AG - MA gradbeništvo d.o.o. (9.816171943337283)
3375374000 :: PARK INVESTICIJE gradnje za trg d.o.o. - v stečaju (9.796399360048895)
5055218000 :: RENOVA ROMAN LOGAR S.P. (9.683938125106799)
```

Method 3: Period Margin Points Cumulative

In the third method we are defining a financial transaction as a base relation between two entities (public sector entity and business entity). Based on this, we detect relation periods (when relation started or ended) and accumulate starting/ending periods on a timeline. Based on cumulative relation period extremes, we identify deviations and list entities as part of identified extremes.

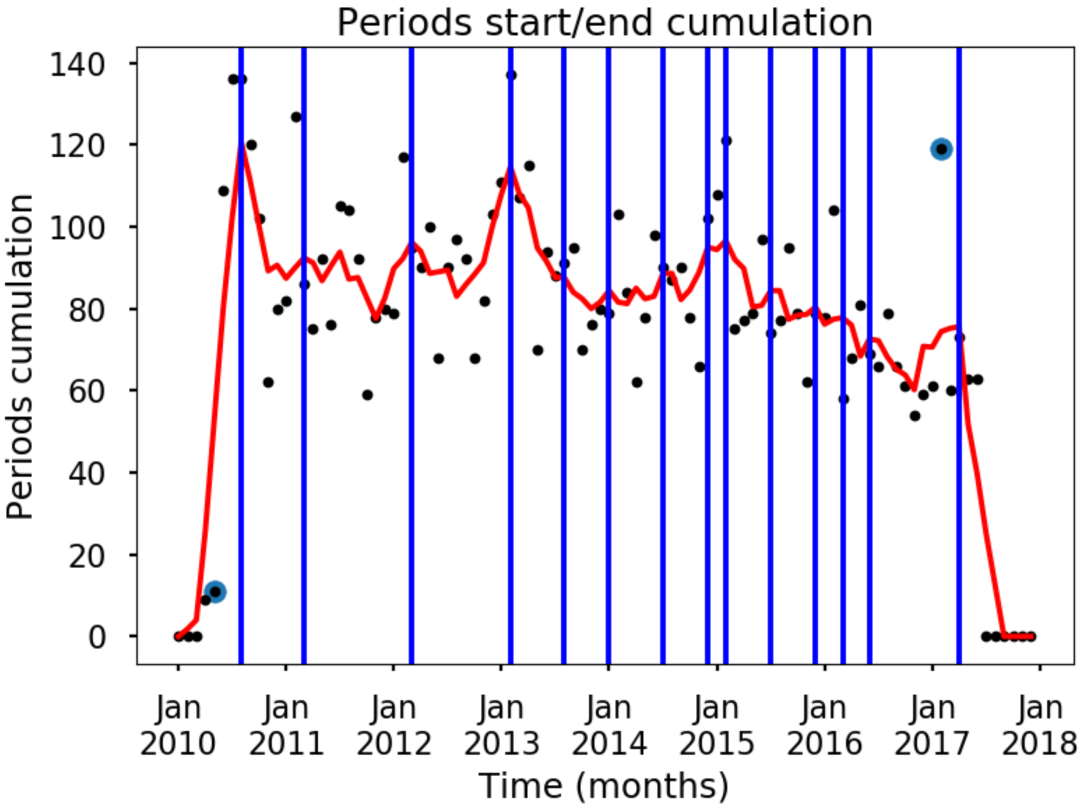
The graphical representation of detection method is the following:



The upper part of the figure represents the sample data. We can see the transactions between public entity and private company started to emerge in October 2014 and stopped in November 2015, and in April 2015 there was a small interruption.

The method detects two deviation points, one when transactions started and the other where transactions ended.

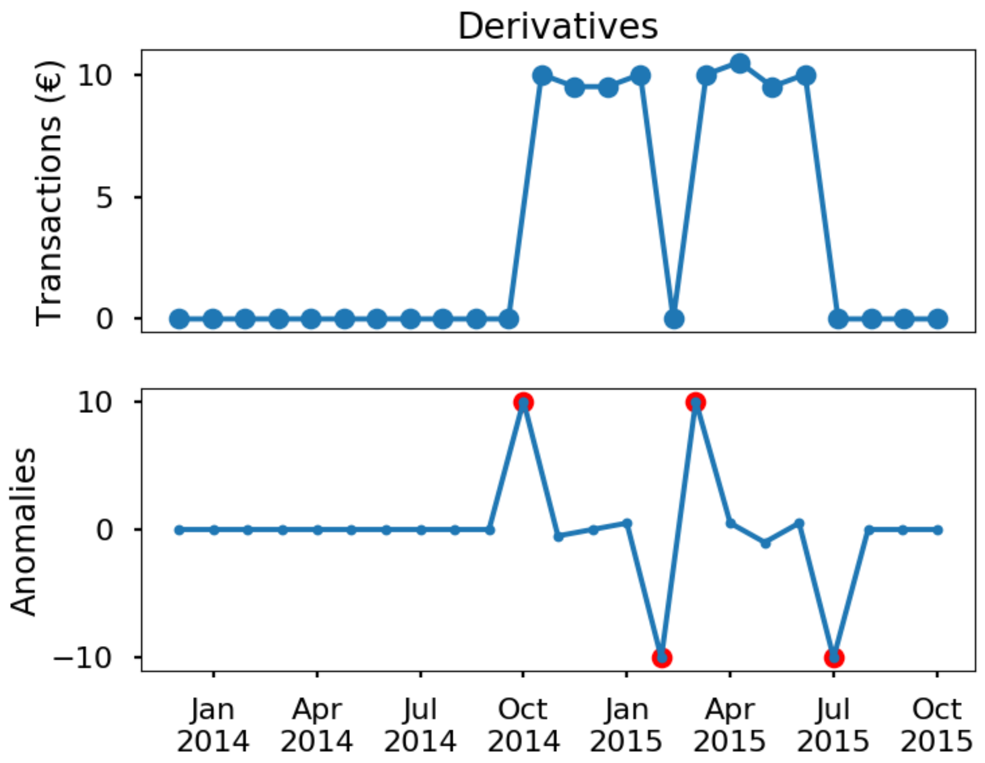
All detected anomalies across all pairs of transactions between public institutions and private companies are then grouped together and presented on a timeline. Here is the example of a timeline of detected deviations in transactions from budget users to private companies registered to build construction:



Method 4: Local Extremes Detection

Similarly to Method 3, Method 4 analyses the biggest changes within two entities and transactional relation in a given period. If a change is identified as an anomaly it is added to the cumulative anomaly graph. Once cumulative anomaly graph is defined (based on all transactions jumps between the two entities) extremes are identified – and anomalous companies identified.

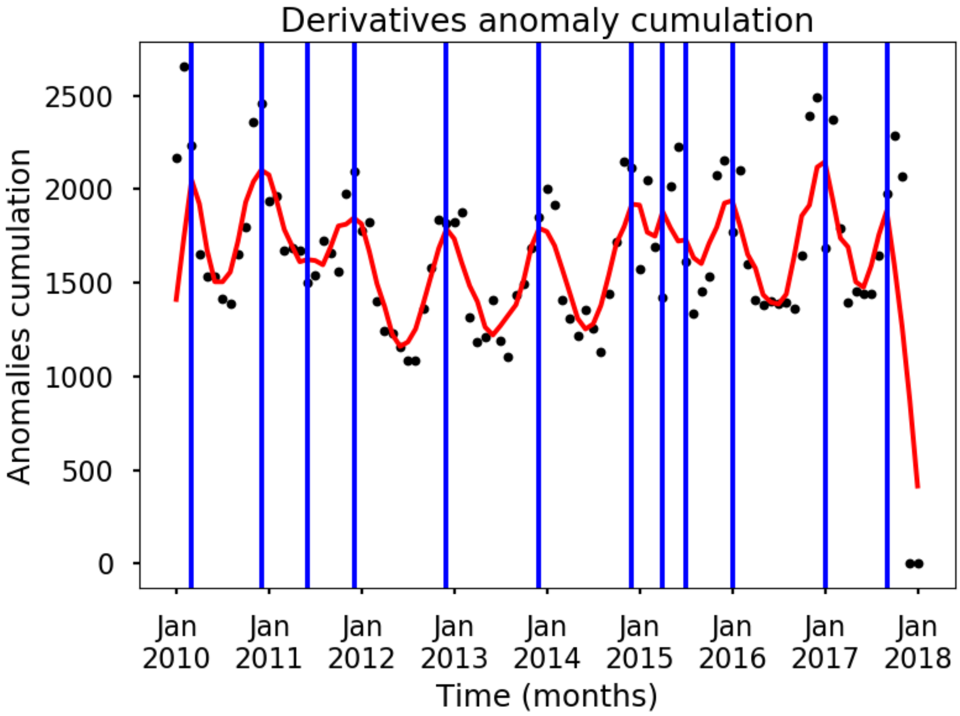
The purpose of the method is to identify the companies manifesting biggest changes in transaction relations.



The upper part of the figure represents the sample data. We can see the transactions between the public entity and the private company started to emerge in October 2014 and stopped in November 2015, and in April 2015 there was a small interruption.

The method will detect four deviation points and we can also see the volume and direction of anomaly.

All detected anomalies across all pairs of transactions between the public institutions and the private companies are then grouped together and presented on a timeline.

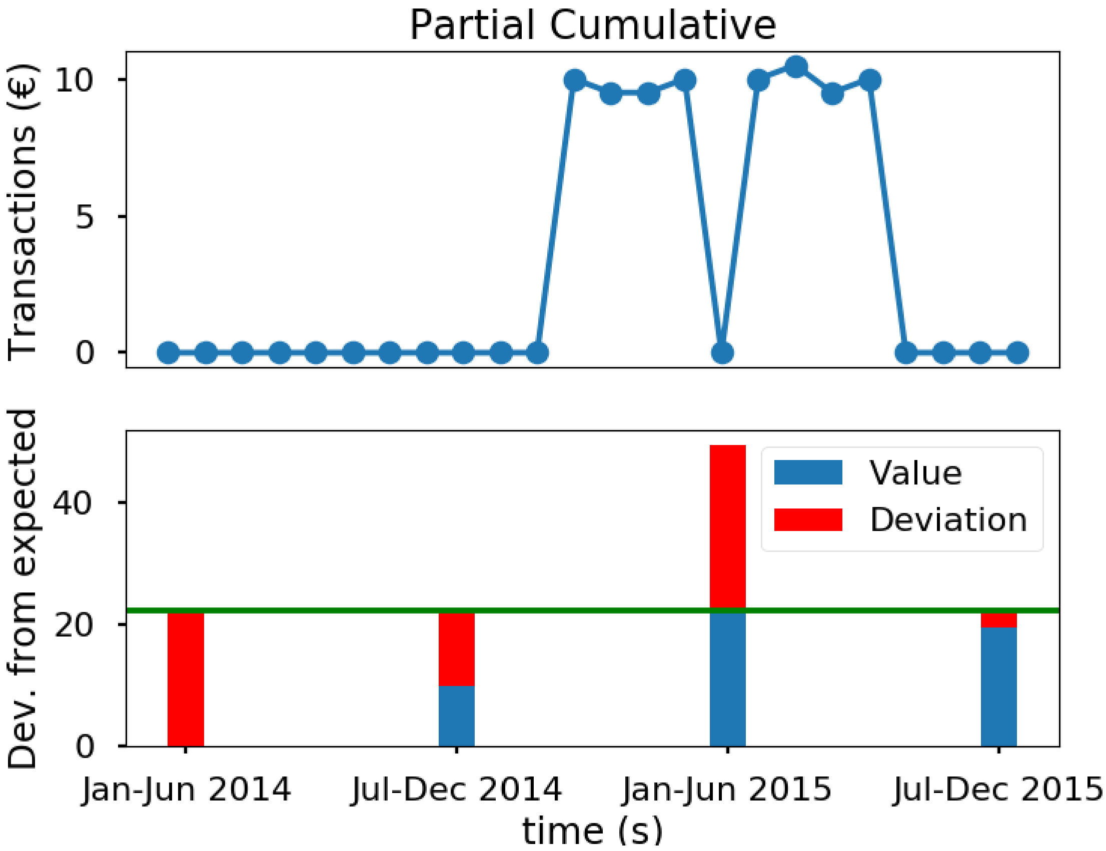


The figure presents grouped anomalies (on a timeline) in transactions from the public sector and the private companies registered to computer programming and IT consulting. We can clearly see the pattern, when anomalies typically occur at the end of the year, which is also the end of the fiscal year. At that time the public sector institutions in Slovenia usually want to spend all their budgets; if not, they are required to return the money to the central budget.

Method 5: Time Periods Deviations

Method 5's approach first defines transaction sums for all related entities and normalizes sums with the total transactions sum. In such way Method 5 defines a comparison baseline. Then, it takes transactions between entities and sums them into a predefined number of periods. For each period partial sums weights are compared to the baseline weights and anomalies are identified. The more anomalous a company behaves the higher it ranks on the anomalous list.

The purpose of the method is to identify the biggest changes within the series of accumulated periods.



The upper part of the figure represents the sample data. We can see the transactions between the public entity and the private company started to emerge in October 2014 and stopped in November 2015, and in April 2015 there was small interruption.

The method will first compute the average amount of transaction between the public institution and the private company. This will detect deviation from this average in the different time periods.

3 Used data

We have tested developed anomaly detection methods with Slovenian data. The data we have used are described below.

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 through this URL:

https://www.ajpes.si/Doc/AJPES/Za_razvijalce/Struktura_podatkov_PRS_042016.doc

The 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 the PostgreSQL is available on Github: https://github.com/zejn/mdb2csv.

The database is available via Access to the Public Information Act, which has been a very time - consuming procedure. Ajpes agency is known to be very unresponsive to access to the 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 the public entities and the 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/

The 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. The 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. The 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 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.

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.

4 Application code

The application code for anomaly detection is published on our Github repository:
https://github.com/TBFY/transactions-data-analysers

5 Conclusions and future work in the area of anomaly detection

We have presented the developed methods for anomaly detection in public spending, which is the key approach for automatic real-time monitoring and analysis of public spending data. Our methods are suitable for detecting anomalies in structured data, for instance on financial transactions between two entities.

With our methods, we are trying to detect relevant signals in the data. However, for the final assessment whether detected anomaly has a meaning – if that it is really relevant in terms of misconduct in public spending or in terms of being an example of extremely good practice in public spending. Sometimes expert human experience is needed. Our system is capable of processing tens of millions of transactions and allows detecting a large class of anomalies in the automatic mode or in the exploratory mode (with human-machine interaction). Therefore some of our methods could be used as a helpful tool for experts to help them extract valuable knowledge from a large amount of data which are usually hard to understand for humans.

Our future work will be oriented towards further development and refining already implemented anomaly detection methods.