(Un)certain Anomalies in Income Data

Dr. Ralph Foorthuis

Lead architect UWV Data Services

January 28th 2019

Academiegebouw, Utrecht University

Mini-Symposium on Uncertainty in Data-Driven Systems

(Un)certain Anomalies in Income Data:
Anomaly Detection for Data Quality Analysis

Topics:

A. Dutch Alliance on Income Data and Taxes

- The value chain: “Loonaangifteketen”
- The Polis Administration domain at UWV

B. Data Quality Analysis

- Operational process
- Anomaly detection
- Discussion on uncertainty

Dutch alliance on income data and taxes: Loonaangifteketen

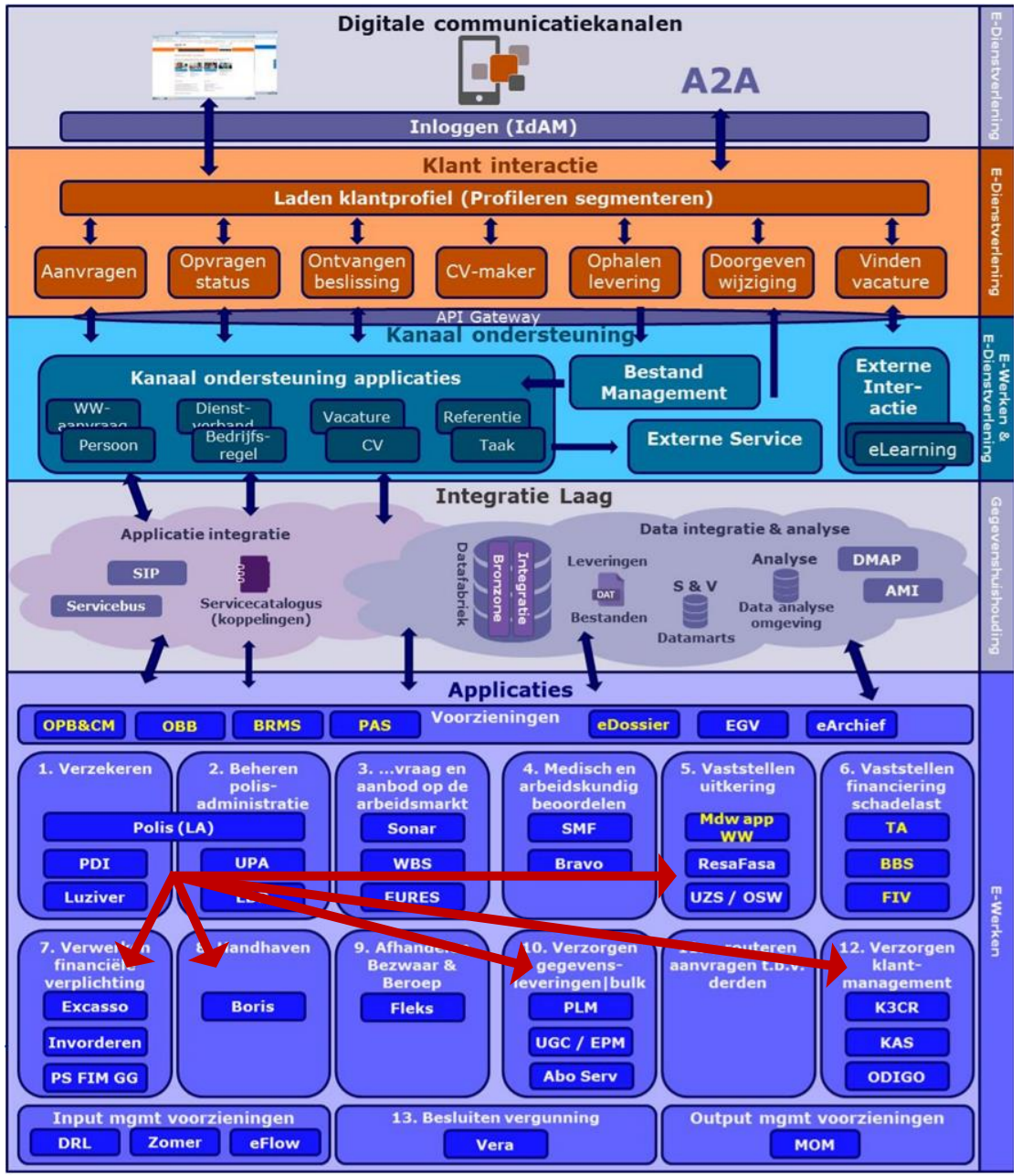
Tax and Customs Office, UWV and Statistics Netherlands

- Some 20 million income relationships (a.o. 'payroll' data / 'salary slips')
- Annual collection of about 150 billion euros of taxes
- Largest source of income for the government of The Netherlands
- Very large ecosystem in terms of stakeholders, IT, processes and data (input & output)
- Goals: Taxes, Data, Lowering of administrative burden
- Won the ASAP Alliance Excellence Award in 2017

UWV Enterprise Achitecture

Data used in:

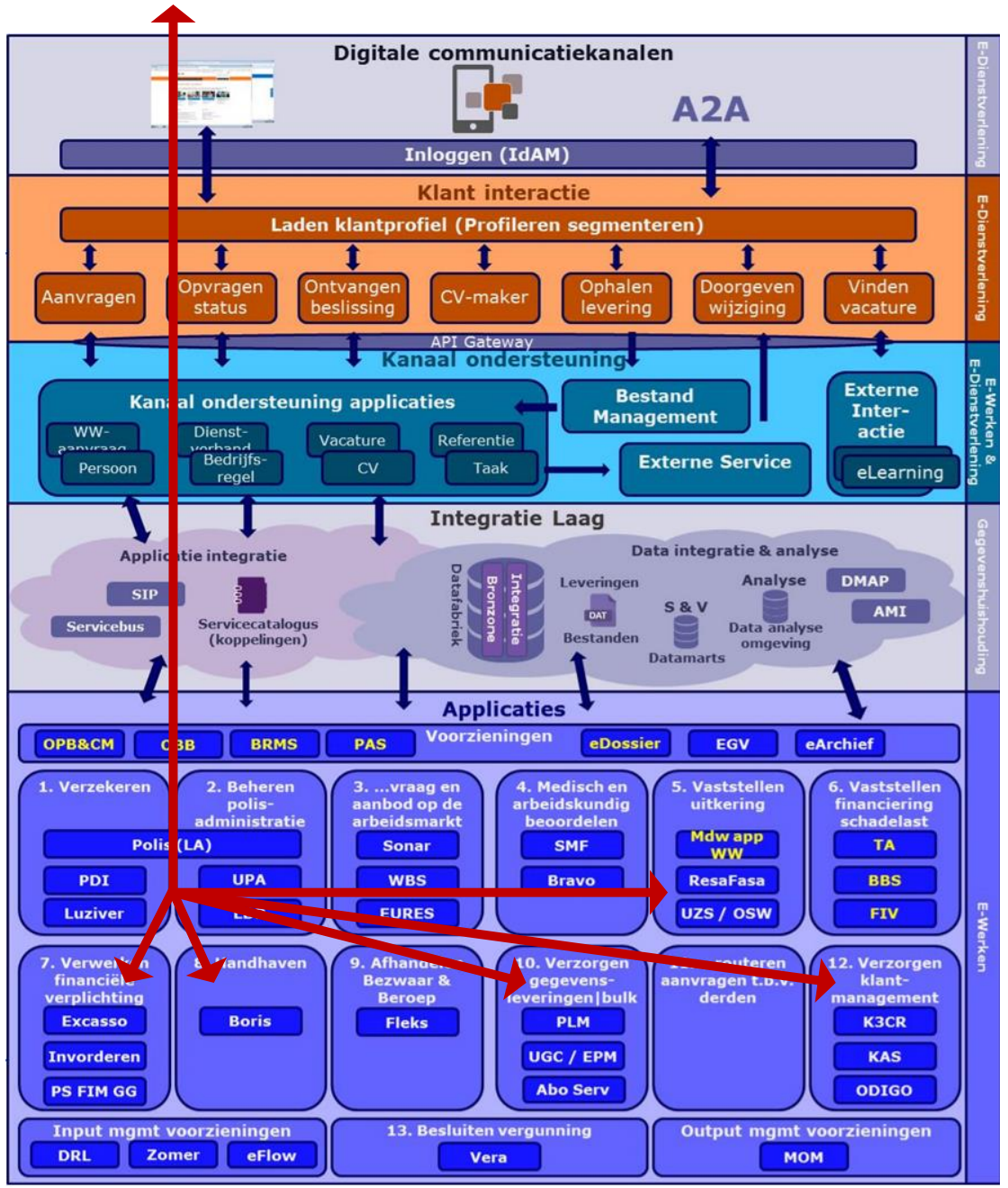
- Internal processes



UWV Enterprise Achitecture

Data used in:

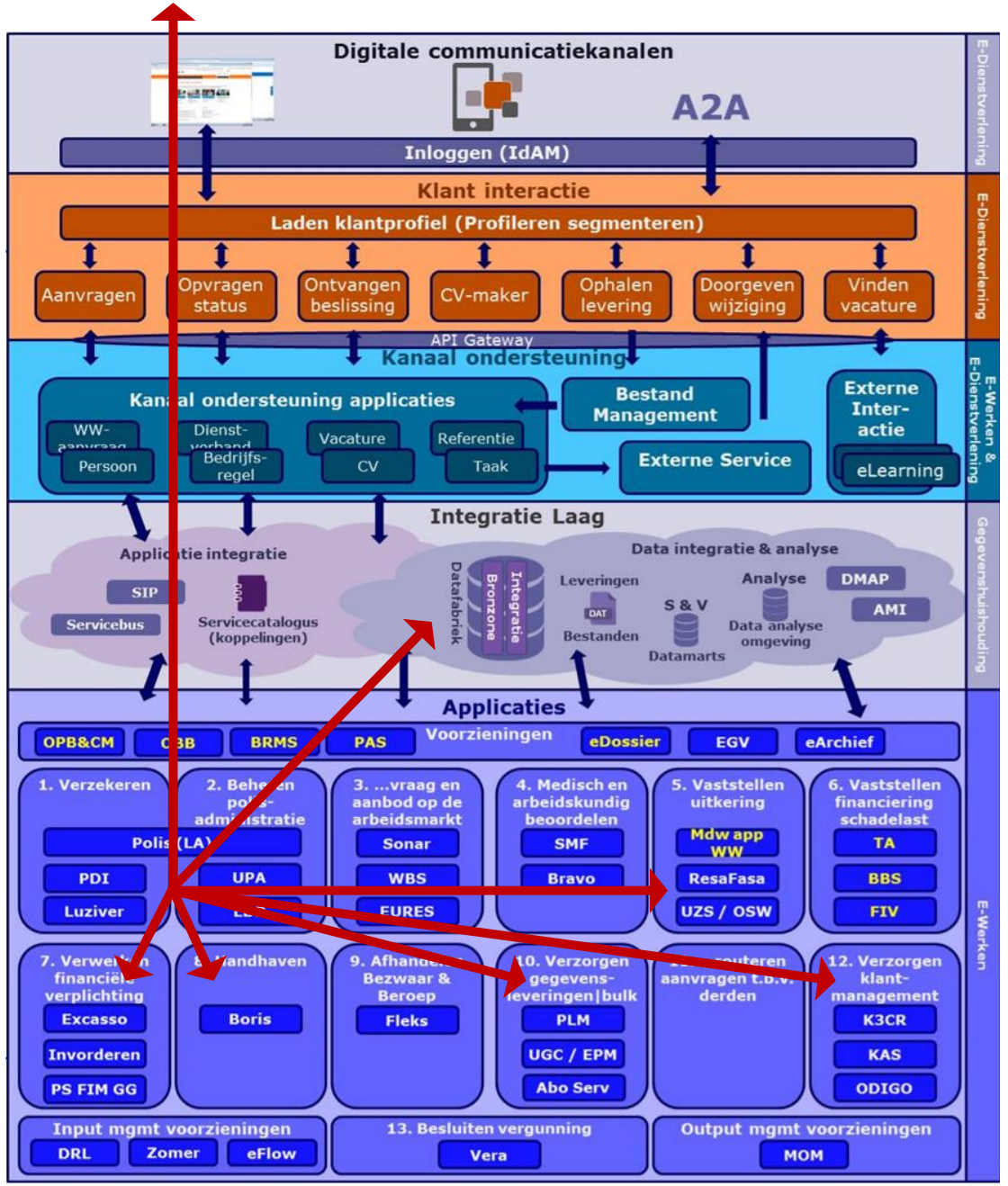
- Internal processes
- External processes (e.g. municipalities, pension funds, CBS, SVB, CAK)



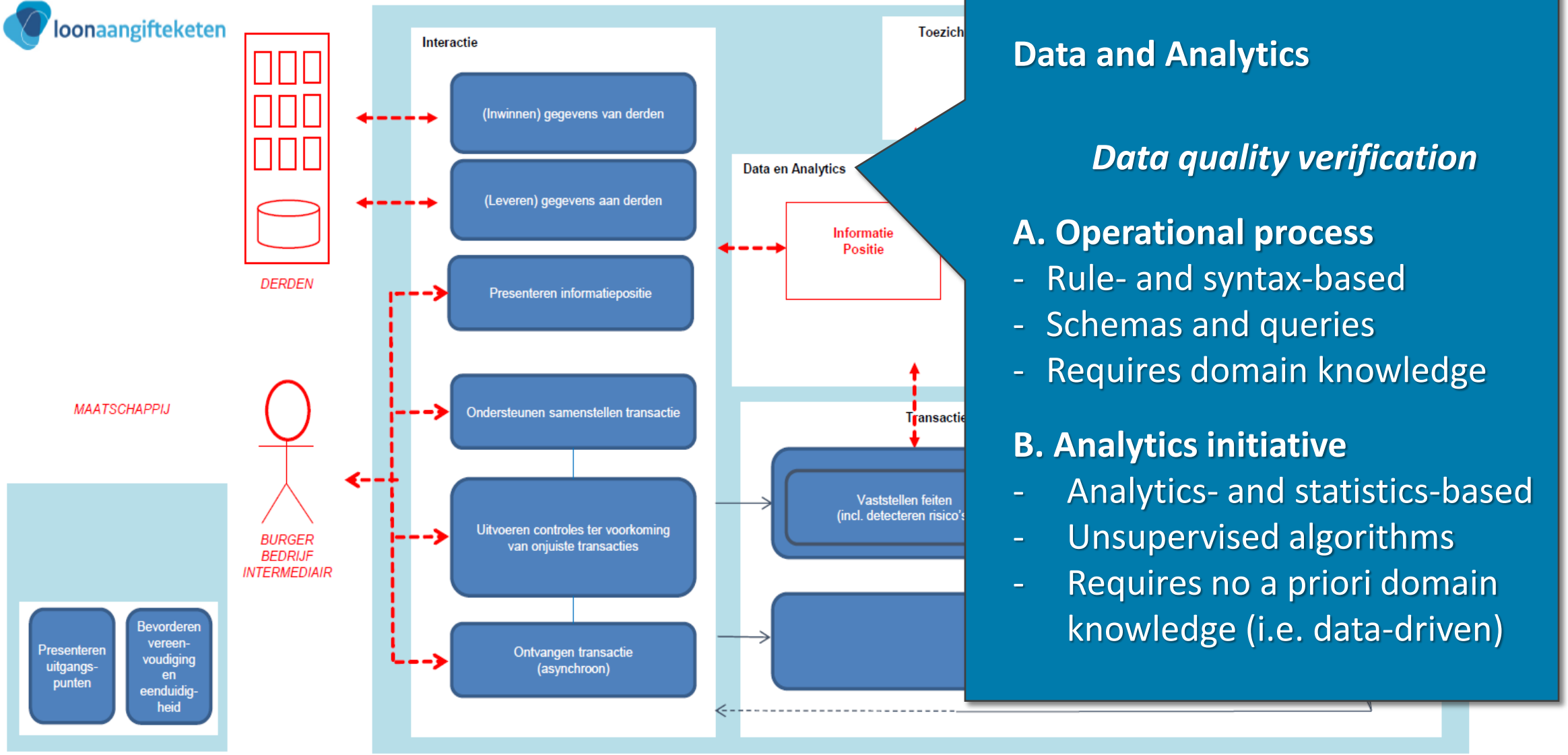
UWV Enterprise Achitecture

Data used in:

- Internal processes
- External processes (e.g. municipalities, pension funds, CBS, SVB, CAK)
- Data analysis (e.g. quality verification, fraud detection)



loonaangifteketen



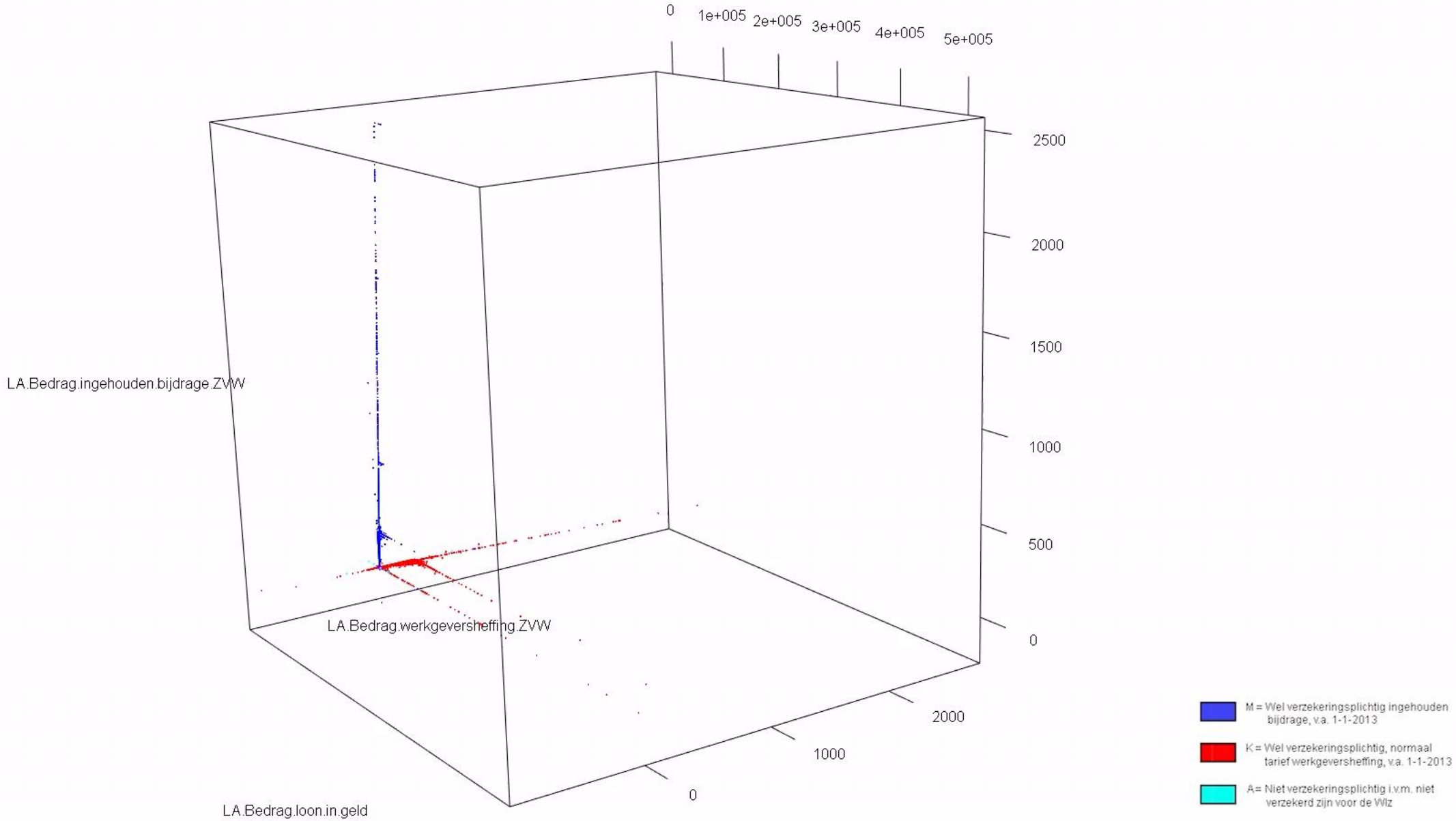
Data and Analytics

Data quality verification

A. Operational process
- Rule- and syntax-based
- Schemas and queries
- Requires domain knowledge

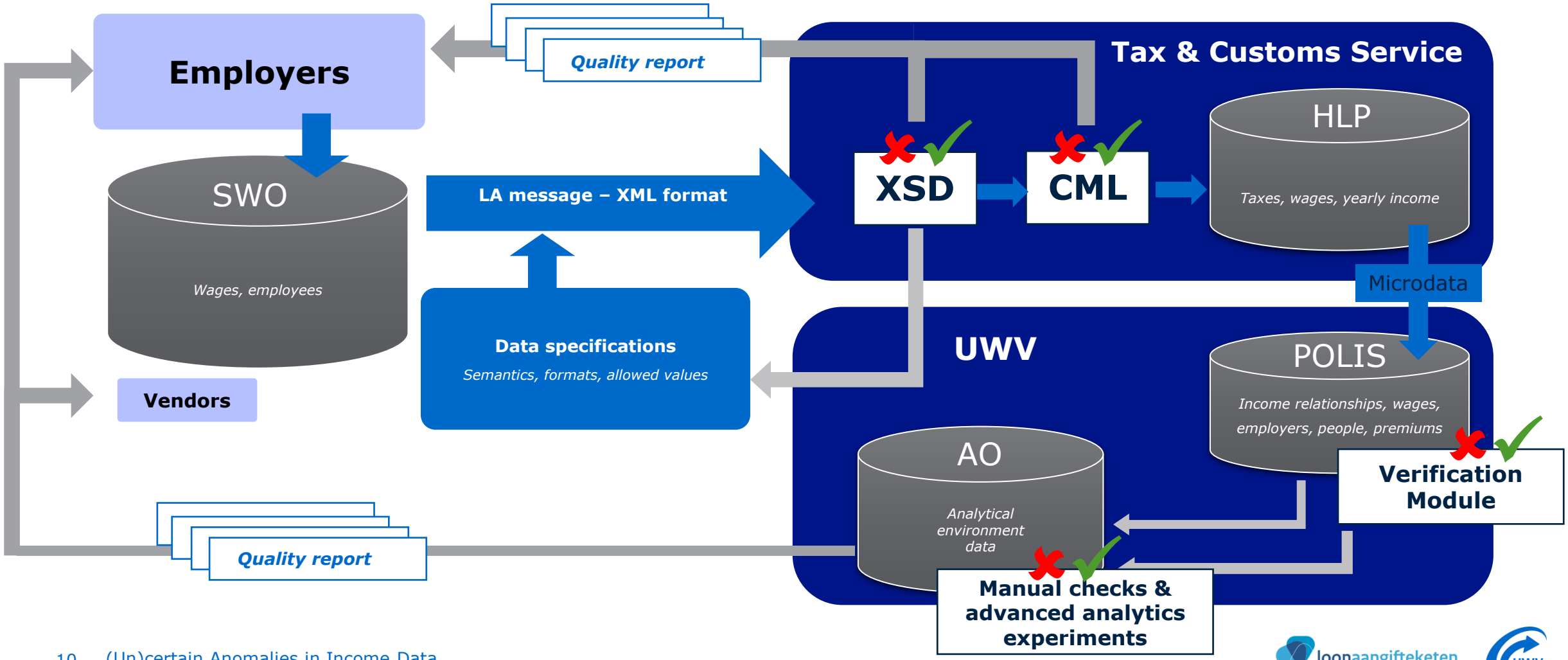
B. Analytics initiative
- Analytics- and statistics-based
- Unsupervised algorithms
- Requires no a priori domain knowledge (i.e. data-driven)

loonaangifteketen uwv


0
1e+005
2e+005
3e+005
4e+005
5e+005
2500
2000
1500
1000
500
0
LA.Bedrag.ingehouden.bijdrage.ZVW
LA.Bedrag.werkgeversheffing.ZVW
2000
1000
0
LA.Bedrag.loon.in.geld
M = Wel verzekeringsplichtig ingehouden bijdrage, v.a. 1-1-2013
K = Wel verzekeringsplichtig, normaal tarief werkgeversheffing, v.a. 1-1-2013
A = Niet verzekeringsplichtig i.v.m. niet verzekerd zijn voor de Wlz
Enz...

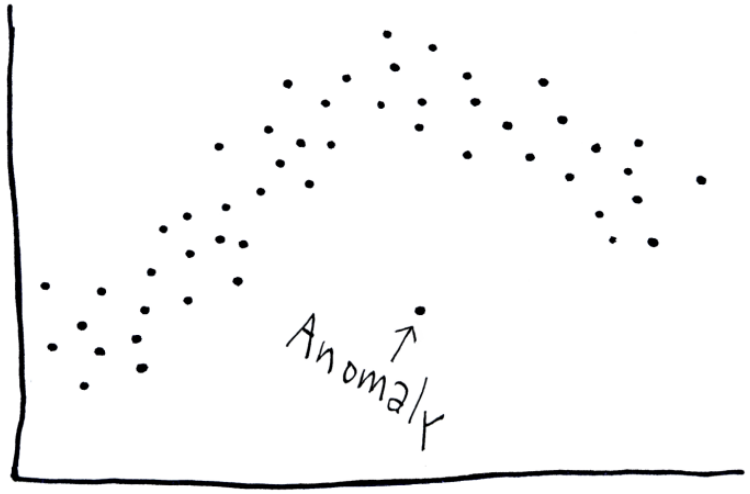
Operational process

Compliance with data quality standards is verified at several locations in the chain, depending on the type of check.



Anomaly detection

- Identifying anomalies (deviant cases) by employing non-trivial algorithms.
- Anomalies are cases that do not fit the general patterns in the dataset.
- There are many potential causes that could explain the anomaly. The deviation may very well point to incorrect data.
- Also known as *outlier detection* and *novelty detection*.
- Application areas:
  - Data quality
  - Fraud detection
  - Error detection and process monitoring
  - Information security
  - Building statistical models



Academic publications

ACCEPTED FOR PRESENTATION AT THE 4TH IEEE INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS (DSAA'2017)

SECODA: Segmentation- and Combination-Based Detection of Anomalies

Ralph Foorthuis
Data Services / CIO Office
UWV, Amsterdam, the Netherlands
ralph.foorthuis@uwv.nl

Abstract — This study introduces SECODA, a novel general-purpose unsupervised non-parametric anomaly detection algorithm for datasets containing continuous and categorical attributes. The method is guaranteed to identify cases with unique or sparse combinations of attribute values. Continuous attributes are discretized repeatedly in order to correctly determine the frequency of such value combinations. The concept of constellations, exponentially increasing weights and discretization cut points, as well as a pruning heuristic are used to detect anomalies with an optimal number of iterations. Moreover, the algorithm has a low memory imprint and its runtime performance scales linearly with the size of the dataset. An evaluation with simulated and real-life datasets shows that this algorithm is able to identify many different types of anomalies, including complex multi-dimensional instances. An evaluation in terms of a data quality use case with a real dataset demonstrates that SECODA can bring relevant and practical value to real-world settings.

Keywords — Anomaly detection; Non-parametric data mining; Unsupervised learning; Mixed data; Data quality; Fraud detection; SECODA; Outlier identification; Discretization; Data visualization

I. Introduction

Anomaly detection (AD) aims at identifying cases that are in some way awkward and do not appear to be part of the general pattern(s) present in the dataset [1, 2, 3, 4, 5, 6]. Such an analysis is often also referred to as *novelty detection* or *outlier identification* [7]. Anomaly detection can be used for various goals, such as fraud detection, data quality analysis, security scanning, process monitoring and data cleansing prior to statistical modelling.

Depending on the specific situation and goals, an anomaly can be taken to mean different things. As [8] put it, "one person's noise is another person's signal". Nonetheless, several types of cases can generally be acknowledged as anomalies. The Theory section presents an overview of these types.

This article presents a novel unsupervised non-parametric anomaly detection algorithm for datasets containing continuous (numerical) and/or categorical attributes. The algorithm in case is SECODA, SEgmentation- and COmbination-based Detection of Anomalies, the core purpose of which is the identification of different types of anomalies. The algorithm is deliberately kept simple for several reasons. First, it is relevant for academia to know that sophisticated AD analysis results can be obtained by relatively simple (and hitherto unexplored) principles and implementations. Second, it makes it possible for practitioners to implement the algorithm on basic platforms, such as machines with relatively little memory and simple DBMS systems that do not offer support for advanced analytics. It also allows for in-database analytics, i.e. analyzing the data in the database itself. This avoids the need to export the data to a separate analytics application, which positively affects time performance and security. The algorithm will therefore be restricted to basic data operations (sort, count, join), control flows (loops) and set-based actions (no point-to-point distances or associations, no complex fitting procedures), so as to show that this can yield sophisticated anomaly detection results.

As part of a real-world evaluation, we will demonstrate how SECODA, and indeed anomaly detection in general, can contribute to improving data quality. In statistics, data quality is evidently important for the analysis process [2, 3]. However, high-quality data is also relevant in broader organizational settings, as it is important for obtaining various kinds of benefits, such as increasing the value of IT systems, enhancing customer service performance, optimizing decision making and improving organizational efficiency [9, 10, 11, 12].

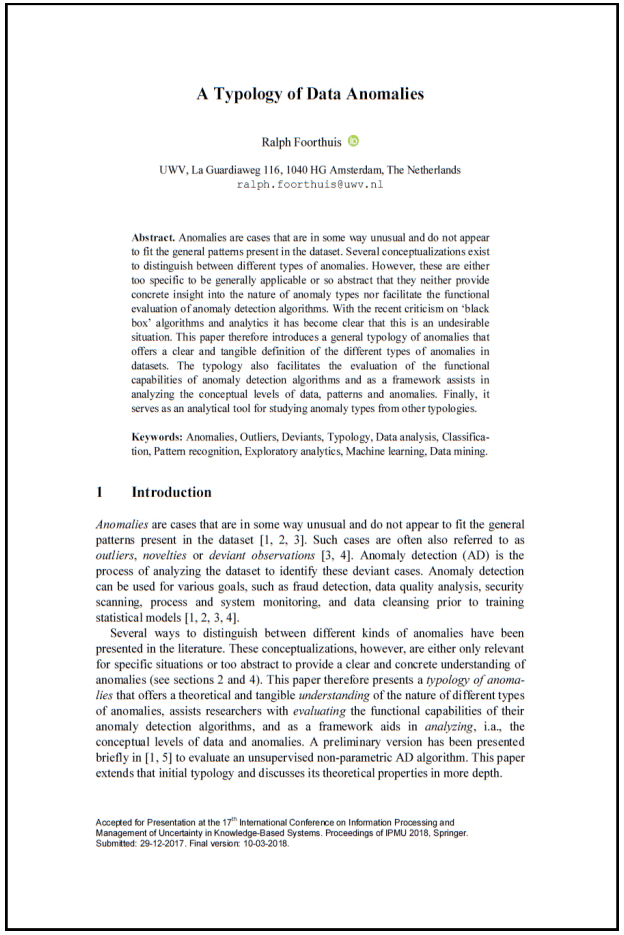
This paper proceeds as follows. The Theory section presents a typology of anomalies, related research and a description of the SECODA algorithm. The Algorithm Evaluation section presents the research approach, results and discussion. The Conclusion summarizes the contributions and discusses further research.

II. Theory

A. Typology of Anomalies

The literature mentions several ways to distinguish between types of anomalies. For example, in sequence or time series analysis, so-called additive, innovational, level shift, and transitory change outliers are often acknowledged [13]. A distinction between weak outliers (noise) and strong outliers (true anomaly) can also be made [1]. In the context of regression analysis it is common to distinguish between outliers, high-leverage points and influential points [2, 3]. More in general, one can differentiate between point, contextual and collective anomalies [14, cf. 15].

The types of anomalies mentioned above are either too specific or too general for the purpose of this study. An alternative typology is therefore presented below. This



A Typology of Data Anomalies

Ralph Foorthuis

UWV, La Guardiaweg 116, 1040 HG Amsterdam, The Netherlands
ralph.foorthuis@uwv.nl

Abstract. Anomalies are cases that are in some way unusual and do not appear to fit the general patterns present in the dataset. Several conceptualizations exist to distinguish between different types of anomalies. However, these are either too specific to be generally applicable or so abstract that they neither provide concrete insight into the nature of anomaly types nor facilitate the functional evaluation of anomaly detection algorithms. With the recent criticism on 'black box' algorithms and analytics it has become clear that this is an undesirable situation. This paper therefore introduces a general typology of anomalies that offers a clear and tangible definition of the different types of anomalies in datasets. The typology also facilitates the evaluation of the functional capabilities of anomaly detection algorithms and as a framework assists in analyzing the conceptual levels of data, patterns and anomalies. Finally, it serves as an analytical tool for studying anomaly types from other typologies.

Keywords: Anomalies, Outliers, Deviants, Typology, Data analysis, Classification, Pattern recognition, Exploratory analytics, Machine learning, Data mining.

1 Introduction

Anomalies are cases that are in some way unusual and do not appear to fit the general patterns present in the dataset [1, 2, 3]. Such cases are often also referred to as *outliers*, *novelties* or *deviant observations* [3, 4]. Anomaly detection (AD) is the process of analyzing the dataset to identify these deviant cases. Anomaly detection can be used for various goals, such as fraud detection, data quality analysis, security scanning, process and system monitoring, and data cleansing prior to training statistical models [1, 2, 3, 4].

Several ways to distinguish between different kinds of anomalies have been presented in the literature. These conceptualizations, however, are either only relevant for specific situations or too abstract to provide a clear and concrete understanding of anomalies (see sections 2 and 4). This paper therefore presents a *typology of anomalies* that offers a theoretical and tangible *understanding* of the nature of different types of anomalies, assists researchers with *evaluating* the functional capabilities of their anomaly detection algorithms, and as a framework aids in *analyzing*, i.a., the conceptual levels of data and anomalies. A preliminary version has been presented briefly in [1, 5] to evaluate an unsupervised non-parametric AD algorithm. This paper extends that initial typology and discusses its theoretical properties in more depth.

Accepted for Presentation at the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Proceedings of IPMU 2018, Springer.
Submitted: 29-12-2017. Final version: 10-03-2018.

The Impact of Discretization Method on the Detection of Six Types of Anomalies in Datasets

Ralph Foorthuis

UWV, La Guardiaweg 116, 1040 HG Amsterdam, The Netherlands
ralph.foorthuis@uwv.nl

Abstract. Anomaly detection is the process of identifying cases, or groups of cases, that are in some way unusual and do not fit the general patterns present in the dataset. Numerous algorithms use discretization of numerical data in their detection processes. This study investigates the effect of the discretization method on the unsupervised detection of each of the six anomaly types acknowledged in a recent typology of data anomalies. To this end, experiments are conducted with various datasets and SECODA, a general-purpose algorithm for unsupervised non-parametric anomaly detection in datasets with numerical and categorical attributes. This algorithm employs discretization of continuous attributes, exponentially increasing weights and discretization cut points, and a pruning heuristic to detect anomalies with an optimal number of iterations. The results demonstrate that standard SECODA can detect all six types, but that different discretization methods favor the discovery of certain anomaly types. The main findings also hold for other detection techniques using discretization.

Keywords: Anomaly detection · Outlier detection · Deviants · SECODA · Data mining · Typology · Discretization · Binning · Classification · Anomaly types

1 Introduction

Anomaly detection (AD) is the process of identifying cases, or groups of cases, that are in some way unusual and do not fit the general patterns present in the dataset [1, 2, 3]. The detection of *anomalies*, which are often also referred to as outliers, deviants or novelties, is a major research topic in the overlapping disciplines of artificial intelligence [4, 5, 6], data mining [7, 8, 9] and statistics [10, 11, 12] . It is not merely of interest for academia, however, as it is also of significant value in industrial practice nowadays [13, 14, 36]. Anomaly detection can be used for discovering fraud, data quality issues, security threats, process and system failures, and deviating data points that hamper model training.

Many techniques for detecting anomalies have been devised throughout the years. The field of statistics traditionally focused mainly on parametric methods for discovering univariate outliers in each attribute (variable) separately [cf. 1, 12, 15]. Distance- and density-based techniques were consequently developed, allowing for non-parametric multidimensional data mining [16, 17, 18]. Another group of methods comprises complex non-parametric models, such as one-class support vector ma-

Accepted for Presentation at the 30th Benelux Conference on Artificial Intelligence (BNAIC 2018), Den Bosch, the Netherlands.

loonaangifteketen

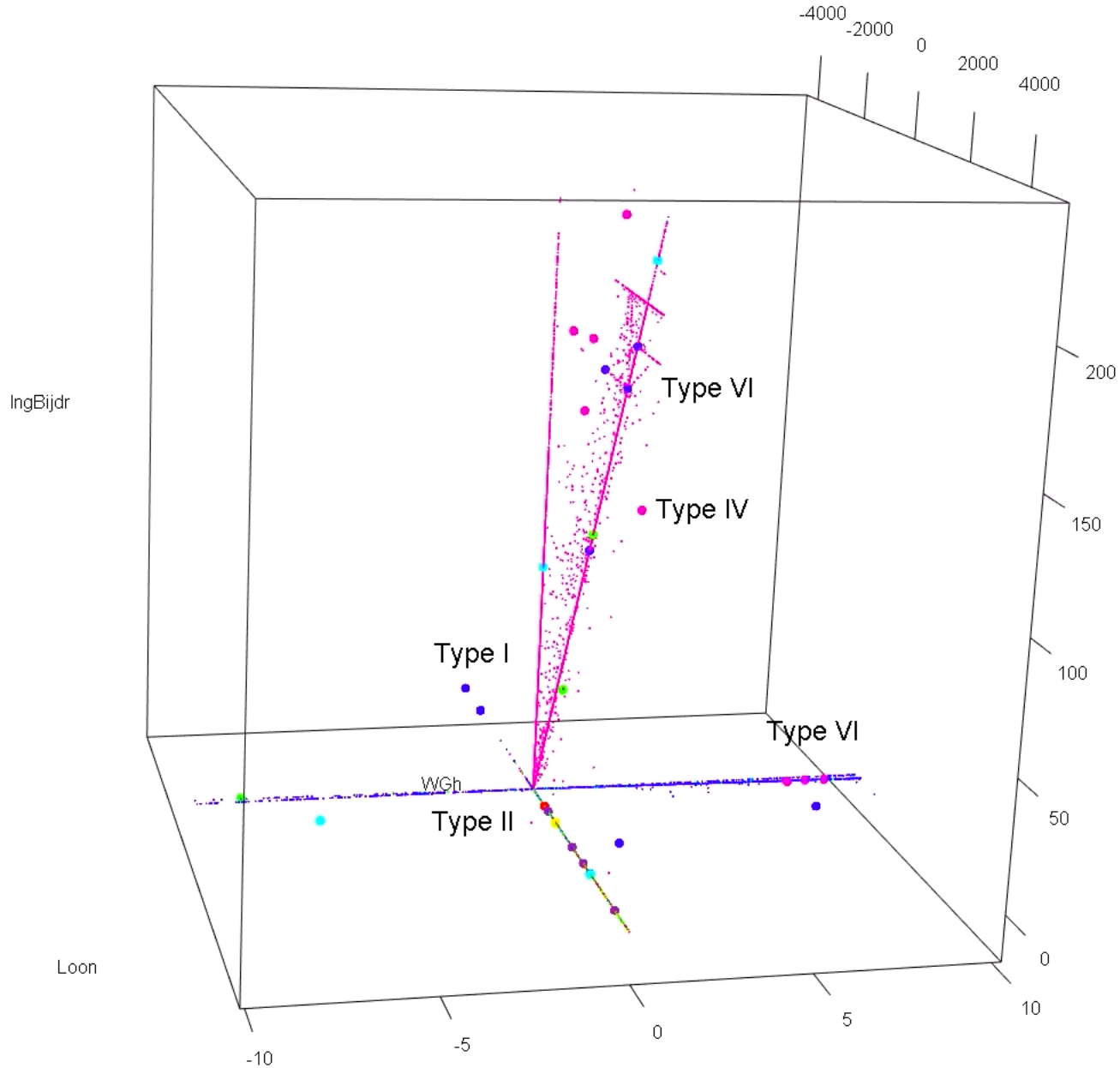
uwv

Typology of anomalies

Cardinality of Relationship		Types of data: Continuous attributes	Types of data: Categorical attributes	Types of data: Mixed attributes
	Univariate Described by individual attributes (independence)	Type I Extreme value anomaly	Type II Rare class anomaly	Type III Simple mixed data anomaly
	Multivariate Described by multi-dimensionality (dependence)	Type IV Multidimensional numerical anomaly	Type V Multidimensional rare class anomaly	Type VI Multidimensional mixed data anomaly

From: Foorthuis (2018), 'A Typology of Data Anomalies', IPMU 2018.

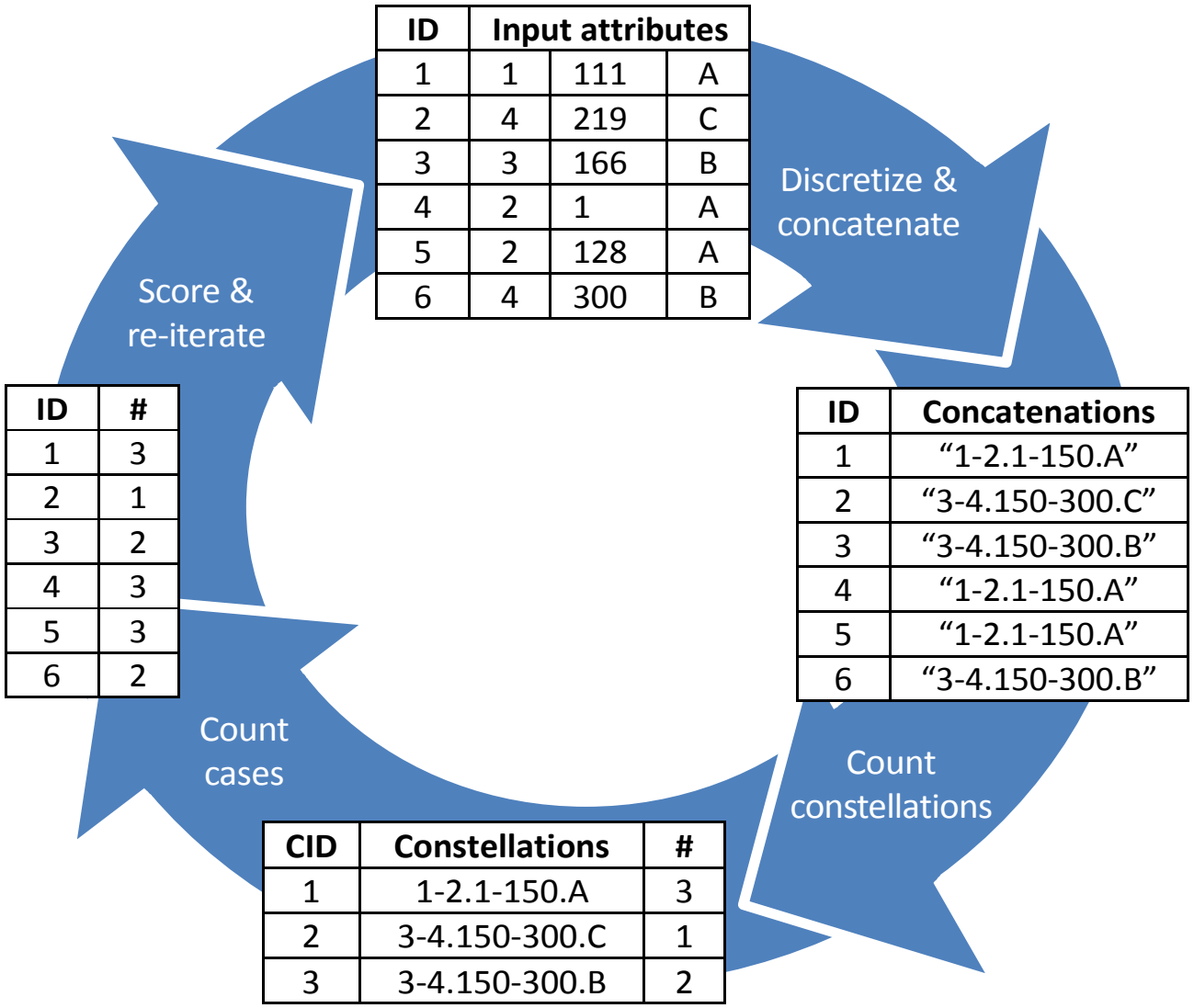
loonaangifteketen uwv


lngBijdr
Loon
WGh
Type VI
Type IV
Type I
Type VI
Type II
-4000
-2000
0
2000
4000
200
150
100
50
0
10
5
0
-5
-10

SECODA

- SECODA is a novel generally applicable algorithm for anomaly detection
- An *algorithm*: a series of instructions (processed by a computer) that leads from a starting state to a target state
- SECODA can be characterized as:
  - Unsupervised: Analyzes data that have not been seen and labelled by humans before
  - Non-parametric: Does not have any assumptions regarding the distribution of the data
  - Can analyze mixed data (numerical & categorical)
  - Multivariate analysis (takes relationships into account)

The SECODA algorithm



ID	Input attributes		
1	1	111	A
2	4	219	C
3	3	166	B
4	2	1	A
5	2	128	A
6	4	300	B

ID	Concatenations
1	“1-2.1-150.A”
2	“3-4.150-300.C”
3	“3-4.150-300.B”
4	“1-2.1-150.A”
5	“1-2.1-150.A”
6	“3-4.150-300.B”

CID	Constellations	#
1	1-2.1-150.A	3
2	3-4.150-300.C	1
3	3-4.150-300.B	2

ID	#
1	3
2	1
3	2
4	3
5	3
6	2

The SECODA algorithm

Algorithm: SECODA
Inputs: $\boldsymbol{D}_0$, the original matrix with n cases and p attributes.
Output: $\boldsymbol{aas}_i$, a vector of average anomaly scores after the last iteration for all cases in $\boldsymbol{D}_0$, with $aas_{g,i}$ representing the individual score.
Key local vars: b, the number of discretization bins (arity).
s, used as stop point and for increased binning.
$cf_{g,i}$, the current frequency in iteration i of the constellation to which case g belongs.

begin
$i \leftarrow 0$; $b \leftarrow 2$; $s \leftarrow 1$; *continue* ← TRUE # Set initial values
while *continue* = TRUE **do**
$i \leftarrow i + 1$
$\boldsymbol{D'} \leftarrow \boldsymbol{D}_i$ with numerical attributes discretized into b equiwidth bins
$cf_{g,i} \leftarrow$ ConstellationFrequencyPerCase($\boldsymbol{D'}$)
if $i > 1$ # Calculate average anomaly scores for cases in $\boldsymbol{D}_i$
$aas_{g,i} \leftarrow \frac{1}{2}(aas_{g,i-1} + cf_{g,i})$
else # If it's the first iteration, put in the frequency
$aas_{g,i} \leftarrow cf_{g,i}$
end if
if $i \leq 10$ # Iteration management
$s \leftarrow s + 0.1$
$b \leftarrow b + 1$
else # Take larger steps and prune cases in higher iterations
$s \leftarrow s + 1$
$b \leftarrow b + (s - 2)$
Add to $\boldsymbol{aasp}_i$ the anomaly scores of the 5% most normal cases that are to be pruned away:
$\boldsymbol{p} \leftarrow$ subset of $\boldsymbol{aas}_i$, with each $aas_{g,i} \geq 0.95$ quantile value
$\boldsymbol{aasp}_i \leftarrow \boldsymbol{aasp}_{i-1} \cup \boldsymbol{p}$
Prune away high-frequency (normal) cases for next iteration:
$\boldsymbol{D}_{i+1} \leftarrow$ subset of $\boldsymbol{D}_i$, with each case such that its $aas_{g,i} < 0.95$ quantile value
end if
$\boldsymbol{Q} \leftarrow$ Subset of $\boldsymbol{D}_i$, with each case such that its $aas_{g,i} \leq s$
if (noc($\boldsymbol{Q}$) / noc($\boldsymbol{D}_0$)) > 0.003 # Verify fraction of identified anomalies
continue ← FALSE # No new iteration (process has converged)
end if
end while
$\boldsymbol{aas}_i \leftarrow \boldsymbol{aas}_i \cup \boldsymbol{aasp}_{i-1}$ # Combine average anomaly scores from latest iteration with scores from cases that have been pruned previously
return $\boldsymbol{aas}_i$ # Return full anomaly score vector as the end result
end

See the DSAA 2017 paper and open source R code for precise implementation and detailed comments.

(See 'SECODA resources for R' at www.foorthuis.nl/)

Algorithm: ConstellationFrequencyPerCase
Inputs: $\boldsymbol{D'}$, containing p (categorical and discretized numerical) attributes and a total of n cases, with $n \leq$ noc($\boldsymbol{D}_0$).
Output: $\boldsymbol{cf}_i$, a vector with for each case $cf_{g,i}$ the frequency of the constellation to which the case belongs in the current iteration.
begin
# Concatenate each case's attribute values in this iteration (i.e. determine the constellations):
$cc_{g,i} \leftarrow d'_{g,1,i} \oplus d'_{g,2,i} \oplus \ldots \oplus d'_{g,p,i}$
Determine the frequency of distinct constellations in this iteration (with k identifying the constellations):
$ccf_{k,i} \leftarrow$ The number of cases per constellation
Determine the frequency of each case, using the frequencies of their constellations (i.e. inner join $\boldsymbol{cc}_i$ and $\boldsymbol{ccf}_i$ on k):
$cf_{g,i} \leftarrow$ The frequency from $ccf_{k,i}$ for each case's corresponding constellation
return $\boldsymbol{cf}_i$ # Return each case's current frequency $cf_{g,i}$ as the elements of a vector
end

loonaangifteketen uwv

Practitioner case at UWV

PLM engine

- Parameterized data extraction module (“Parametriseerbare Lever Module”).
- Two representative (anonymized) samples from the Polis Administration.
- A data point or case in these datasets represents an income relationship.

Analysis ignoring any domain knowledge

- The 30 most extreme anomalies were scrutinized.
- These can indeed all be considered data anomalies (see 4D plot).
- All possible anomaly types were present.

Analysis taking into account domain knowledge

- Most anomalies could be explained as normal or could not be proven erroneous.
- However, several identified anomalies (positioned both isolated and in-cloud) did prove to be indicative of previously unknown data quality problems. Interestingly, the anomalies represented a broader issue and turned out to be data export complexities rather than errors in the stored data.
- These results resulted in changes to improve the operational data delivery software.

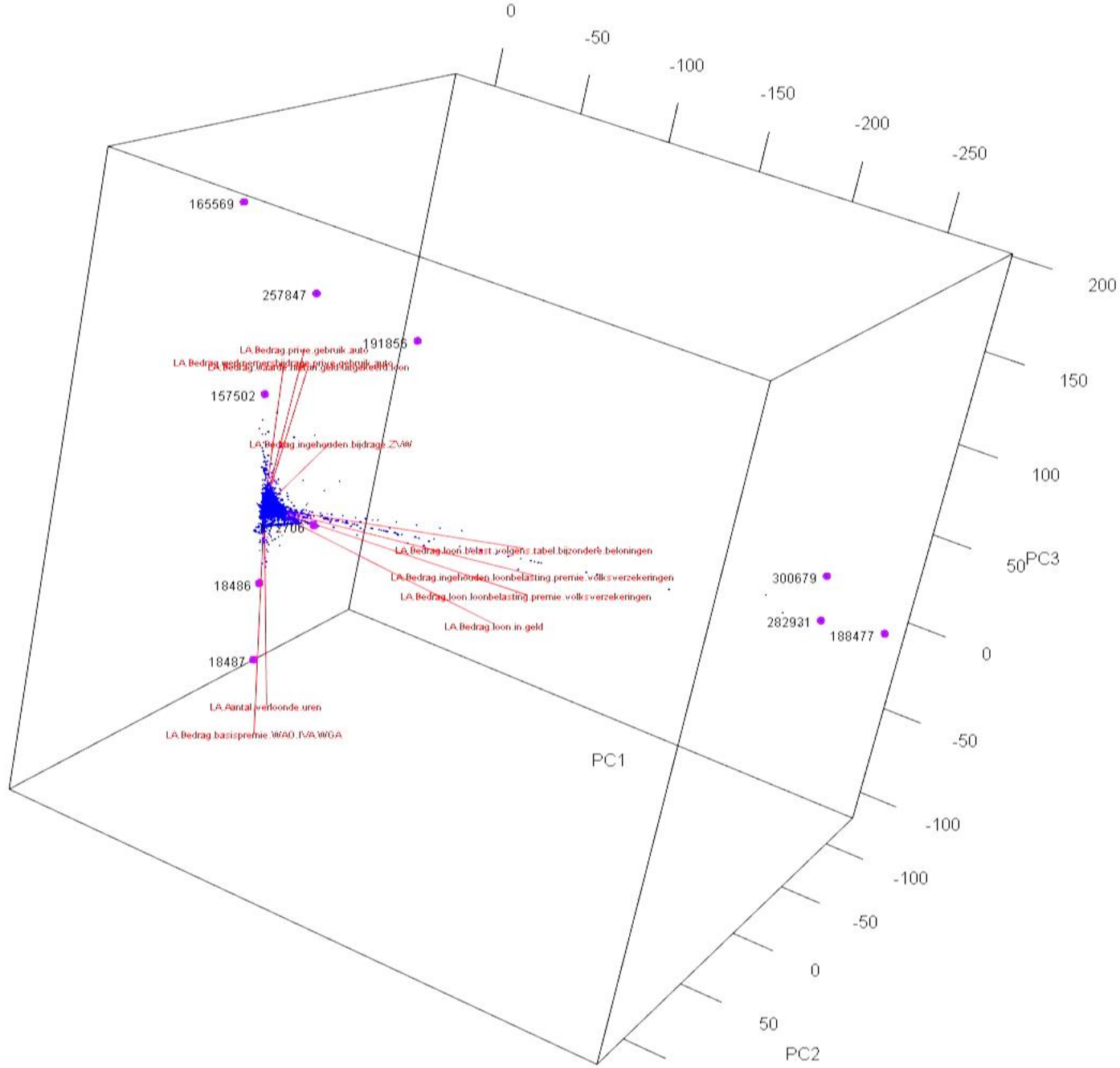
Findings on unsupervised anomaly detection

Academic findings

- The typology offers a theoretically and empirically justified overview of anomaly types.
- SECODA is able to identify all acknowledged anomaly types (for independent data).
- See the publications for details.

Practical findings

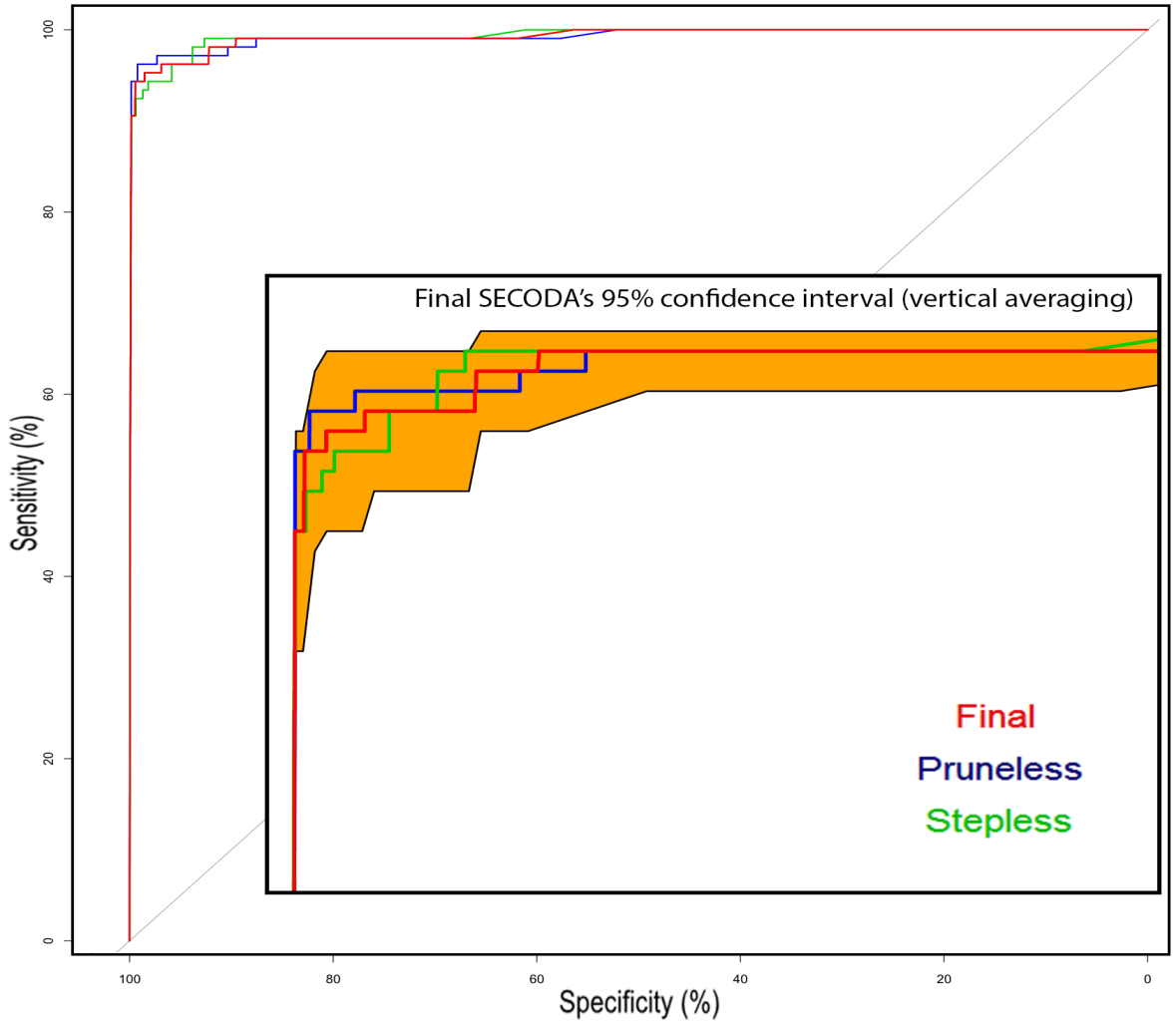
- Anomaly detection (AD) is useful for finding quality issues in real-world administrative data.
- There was a substantial overlap between anomalies found by AD and the cases that were identified by manual quality queries.
- Especially valuable for exploratory analysis, in settings in which one lacks deep domain knowledge, and when a predominantly data-driven analysis is desired (e.g. to complement a rule-based data quality approach).
- Anomalies do not necessarily represent erroneous data.
- For interpretation it is important to have functional knowledge of the AD algorithms.
- Data visualization is also essential for understanding the anomalies.
- High-dimensionality remains a problem.


0
-50
-100
-150
-200
-250
165569
257847
191856
LA.Bedrag.prive.gebruik.auto
LA.Bedrag.werknemersbijdrage.prive.gebruik.auto
157502
LA.Bedrag.ingehouden.bijdrage.ZVW
18486
LA.Bedrag.loon.belast.volgens.tabel.bijzondere.beloningen
LA.Bedrag.ingehouden.loonbelasting.premie.volksverzekeringen
LA.Bedrag.loon.loonbelasting.premie.volksverzekeringen
LA.Bedrag.loon.in.geld
18487
LA.Aantal.verloonde.uren
LA.Bedrag.basispremie.WAO.IVA.WGA
PC1
300679
282931
188477
200
150
100
50
PC3
0
-50
-100
-100
-50
0
50
PC2

SECODA evaluation and uncertainty

We can discuss uncertainty in terms of ROC, PRC, full and partial AUC, p-values, confidence intervals, confidence bands, Sensitivity, Specificity, Precision, Accuracy, F1 measures, Matthews CC, Cohen's Kappa, etc.

But see the DSAA publication for that. Let's focus on other aspects.

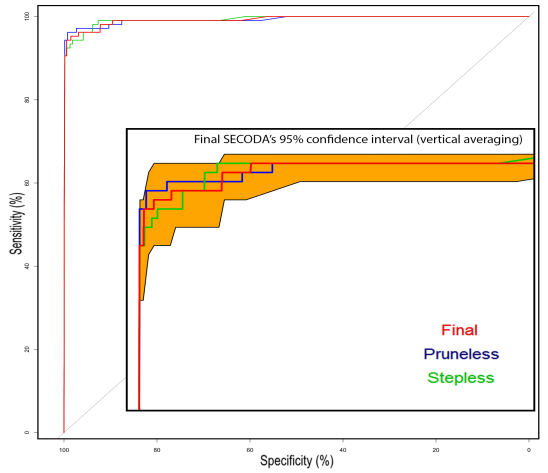


Discussion

Anomaly detection research generally ignores the different anomaly types. Often nothing is known with regard to these types and characteristics of the anomalies. (The dataset could thus be called a ‘black set’.) However, if you don’t know the dataset you are using for evaluating your algorithm, then the results are hardly convincing. So ensure you know your data and especially *which types of anomalies are present* in order to obtain insights in how the evaluated algorithm performs.

Moreover, in order to improve understanding of the data and the algorithm – and thus to decrease uncertainty – the analysis or evaluation should also be done *per anomaly type*.
This provides deeper insights. For example, it may show that certain types can be detected better and with less statistical uncertainty than other types.

		Types of data		
		Continuous attributes	**Categorical attributes**	**Mixed attributes**
Cardinality of Relationship	**Univariate** Described by individual attributes (independence)	Type I Extreme value anomaly	Type II Rare class anomaly	Type III Simple mixed data anomaly
	Multivariate Described by multi-dimensionality (dependence)	Type IV Multidimensional numerical anomaly	Type V Multidimensional rare class anomaly	Type VI Multidimensional mixed data anomaly



Discussion

Some thoughts on uncertainty and knowledge in the context of data quality verification…

Syntax- and rule-based quality verification

- Driven by rules, laws, regulations, technical data formats
- The rules here are based on national laws and strict norms regarding the data model
- Able to verify complex consistency demands between attributes
- Rule-based quality verification can yield conclusions that are (more or less) 100% true
- Sounds comforting – and works especially well for compliance purposes
- But… it's also a bit tautological. Not suited for yielding new ideas & knowledge

Unsupervised anomaly detection

- Algorithmic and truly data-driven; no rules needed
- Able to discover complex inconsistencies between attributes
- Uncertain results; anomalies are not guaranteed to represent erroneous or interesting data
- Results are based on the idiosyncrasies of the algorithm (and the given data distribution)
- You cannot simply automate the decision making process that follows the anomaly detection
- But… it helps discover unexpected and surprising phenomena, and facilitates the creation and expansion of your domain knowledge

Questions?

Sources and References

- Aggarwal, C.C.: *Outlier Analysis*. Springer, New York (2013).
- Beauxis-Aussalet, E.M.A.L. (2019). *Statistics and Visualizations for Assessing Class Size Uncertainty*, Phd Thesis, CWI / Utrecht University. ISBN: 978-90-393-7084-1.
- Boddy, R., Smith, G. (2009). *Statistical Methods in Practice: For Scientists and Technologists*. Chichester: John Wiley & Sons.
- Foorthuis, R.M. (2017). *SECODA: Segmentation- and Combination-Based Detection of Anomalies*. Proceedings of the 4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017), Tokyo, Japan, pp. 755-764.
- Foorthuis, R.M. (2017). *Anomaly Detection with SECODA*. Poster Presentation at the 4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017), Tokyo, Japan.
- Foorthuis, R.M. (2018). *The Impact of Discretization Method on the Detection of Six Types of Anomalies in Datasets*. Proceedings of the 30th Benelux Conference on Artificial Intelligence (BNAIC 2018), November 8-9 2018, Den Bosch, the Netherlands.
- Foorthuis, R.M. (2018). *A Typology of Data Anomalies*. Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018), Cádiz, Spain, CCIS 854.
- Foorthuis, R.M. (2018). *The SECODA Algorithm for the Detection of Anomalies in Sets with Mixed Data*. URL: www.foorthuis.nl, accessed 2017/11/18. (Download SECODA R package and data examples here: https://tunguska.home.xs4all.nl/Publications/AD_resources_for_R.htm)
- Foorthuis, R.M. (2018). *Anomaly Detection at the Dutch Alliance on Income Data and Taxes*. URL: www.loonaangifteketen.nl
- Foorthuis, R.M. (2016). *Anomaliedetectie en Patroonherkenning Binnen de Loonaangifteketen*. ICTU Congres, Digitale overheid van de Toekomst, 28 september 2016.
- Foorthuis, R.M. (in press). *All or In-cloud: How the Identification of Six Types of Anomalies is Affected by the Discretization Method*. To Appear in Springer Nature, Communications in Computer and Information Science, CCIS 1021.
- Foorthuis, R.M., Steenbergen, M. van, Brinkkemper, S., Bruls, W. (2016). *A Theory Building Study of Enterprise Architecture Practices and Benefits*. Information Systems Frontiers, Vol. 18, No. 3, pp. 541-564, DOI: 10.1007/s10796-014-9542-1.
- Foorthuis, R.M., Steenbergen, M. van, Mushkudiani, N., Bruls, W., Brinkkemper, S., Bos, R. (2010). *On Course, But Not There Yet: Enterprise Architecture Conformance and Benefits in Systems Development*. Proceedings of the Thirty First International Conference on Information Systems (ICIS 2010), St. Louis, Missouri, USA.
- Foorthuis, R.M. (2012). *Tactics for Internal Compliance: A Literature Review*. Chapter of 'Project Compliance with Enterprise Architecture', PhD thesis, Utrecht University.
- Foorthuis, R.M., Hofman, F., Brinkkemper, S., Bos, R. (2012). *Compliance Assessments of Projects Adhering to Enterprise Architecture*. Journal of Database Management, Vol. 23, No. 2, pp. 44-71.
- Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York.
- Knorr, E.M., Ng, R.T. (1998). *Algorithms for Mining Distance-Based Outliers in Large Datasets*. VLDB-98, Proceedings of the 24th International Conference on Very Large Data Bases, New York.
- Pimentel, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L. (2014). *A Review of Novelty Detection*. Signal Processing, Vol. 99, pp. 215-249.
- Ramaswamy, S., Rastogi, R., Shim, K. (2000). *Efficient Algorithms for Mining Outliers from Large Data Sets*. Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, USA.

loonaangifteketen uwv