A Proposal for Formalization and Definition of Anomalies in Dynamical Systems

Jan Michael Spoor, Jens Weber, and Jivka Ovtcharova

Abstract Although many scientists strongly focus on anomaly detection in different applications and domains, there currently exists no universally accepted definition of anomalies and outliers. Using an approach based on control theory and dynamical systems, as well as a definition for anomalies as described by philosophy of science, the authors propose a generalized framework viewing anomalies as key drivers of progress for a better understanding of the dynamical systems around us. By mathematically defining anomalies and delimiting deviations within expectations from completely unforeseen instances, this paper aims to be a contribution to set up a universally accepted definition of anomalies and outliers.

Keywords: anomaly detection, outlier analysis, dynamical systems

1 Introduction

Anomalies, often interchangeably called outliers [1], are of key interest in explorative data analysis. Therefore, anomaly detection finds application in many different scientific fields, i.e., in social science, economics, engineering, and medical science [2]. In particular, research in these domains regarding databases, data mining, machine learning or statistics focuses strongly on anomaly detection [3]. Despite the wide

Jan Michael Spoor (✉)
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jan.spoor@kit.edu

Jens Weber
Team Digital Factory Sindelfingen, Mercedes-Benz Group AG, Sindelfingen, Germany, e-mail: jens.je.weber@mercedes-benz.com

Jivka Ovtcharova
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jivka.ovtcharova@kit.edu

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_40

range of anomaly detection, there is currently no universally accepted definition of what an outlier or anomaly is [2], and the mathematical definition depends on the selected method to find these anomalies [4].

The authors previously proposed an applied framework to formalize anomalies within the context of control theory and dynamical systems [5]. In this publication, the idea is discussed in more depth, and a generalization of the framework is proposed to extend its application area to more domains since dynamical systems are relevant in engineering and science [6] as well as in management science and economics [7]. Furthermore, the proposed definition of anomalies should also be applicable outside of the context of control theory and aims to be a contribution to set up a universally accepted definition of anomalies and outliers.

When controlling or simulating dynamical systems, a measurement and prediction process is used. Anomalies occur in this process as substantial deviations of a measured system state (an actual value) from an expected system state (a planned value) [5]. Despite simulation and planning effort, these deviations still occur. While some deviations fall within an acceptable range and within the expectations of normal system behavior, other anomalies are completely unforeseen and do not fit the set-up and expectations of the system. Three sequential questions are derived to further investigate the nature of anomalies within dynamical systems:

1. What distinguishes unforeseen system states from regular system behavior?
2. How can unforeseen system states or errors occur despite simulation?
3. How can unforeseen system states be analyzed and transferred to a standard model of a system's behavior?

2 Definition of Anomalies for Dynamical Systems

2.1 Definitions of Anomalies and Outliers

In general, it is assumed that anomalies are somehow visible within the data of the observed systems. This is also clearly stated by the definition of an outlier or anomaly as data points with a substantial deviation from the norm since this requires a normal state of the system and a measurable deviation [8]. Furthermore, the anomaly detection requires existence and knowledge of a normal state, a definition of a deviation, a metric, and a threshold measure of distance. This threshold measure of distance uses the selected metric. All distances between the norm and the data points, which are either above (in case of distance measures) or below (in case of similarity measures) the defined threshold, are assumed to be non-substantial.

Therefore, in addition, the selection of an appropriate metric becomes an important tool to accurately describe an anomaly. Some authors claim that, in a practical application, the selection of a suitable metric might be more important than the algorithm itself. For example, if clusters are clearly separated within the examined dataset in context of the selected metric, clusters will be found independently of

the used method or algorithm [9]. Other authors claim that the selected method for investigating clusters is of importance [10].

To summarize, there is no trivial definition of a normal state, a deviation, and when a deviation might be substantial. Some authors therefore describe the usefulness of an analysis only within the context of the goals of the analysis [11]. Outlier detection becomes more of a technical target than an actual scientific finding of something novel since the novelty is always defined within the technical target of the analysis. Alternatively, the normal model of the data defines an anomaly [1].

This results, for example, in approaches of regression diagnostics to exclude outliers and anomalous data prior to an analysis or to conduct the analysis along the standard model in a more robust way, which is less affected by anomalies [12]. Both approaches result in the maintaining of the normal model using anomalies as if they were less adequate or not at all representative of the data set.

Since anomalies are only relevant within a context, a typology of anomalies within different dataset contexts can be created. Thus, Foorthuis [13] proposes a typology along the following dimensions: types of data (qualitative, quantitative or mixed), anomaly level (atomic or aggregated) and cardinality of relationship (univariate or multivariate). Anomalies are, within this kind of typology, always dependent on the dataset and behave differently along the measured features, which have been classified as relevant for the specific analysis. The anomaly detection becomes a detection of unfitting, surprising values while maintaining the normal model.

2.2 Definition by Philosophy of Science

If the assumptions regarding normal states, deviation, and substantiality are dropped, it is possible to discuss anomalies on a more fundamental level for understanding our surroundings and the observations of them.

To do this, anomalies have to be placed in the historic context of science and research. Since anomaly detection as a discipline of data science is placed within the scientific context [14], anomaly detection can also be analyzed as part of the scientific method and therefore a comparison with the historical understanding of anomalies in the context of science becomes relevant. By definition of Kuhn [15], anomalies play an important role in the scientific discovery of novelties:

> Discovery commences with the awareness of anomaly, i.e., with the recognition that nature has somehow violated the paradigm-induced expectations that govern normal science. It then continues with a (...) exploration of the area of anomaly. And it closes only when the paradigm theory has been adjusted so that the anomalous has become the expected.

This statement describes scientific progress as a stepwise discovery and the placement of anomalies within a normal state by science. The discussed normal state is therefore dictated by current scientific knowledge, which encompasses the predictions of the currently available and widely used models and theories. An anomaly violates the normal state by violating the predictions of these models. The steps of scientific progress are then as follows:

1. Knowledge of the anomaly.
2. Stepwise acknowledgement of observations and conceptual nature of the anomaly.
3. Change of paradigm and methods to include the anomaly in the new models, often under resistance by the scientific community itself.

Therefore, different states of an anomaly exist as follows:

1. The anomaly is completely unknown.
2. The anomaly is neither described nor modeled but was observed.
3. The anomaly is not commonly recognized and placed within the standard model.

The states of anomalies correspond to the initially defined questions in the introduction regarding the delimitation of anomalous states from normal states, the exploration of the causes for anomalies, and the modeling and planning with the now known anomalies. If the states of anomalies are used to describe practical errors in engineering, error states of systems are not anomalies. This is the case because if error states are priorly classified as such, they are therefore already known and described. This corresponds to the idea that outliers or anomalies are created by a different underlying mechanism [16] and therefore imply an unknown system behavior, which needs modeling to better describe the system. In addition, this follows the assumption of a normal state in which anomalies simply derive from a normal model [1] since they are not part of the normal model. Also, this idea relates strongly to the discussion of the relation between novelty and anomaly detection [17].

To follow the definitions by Kuhn [15], science is driven by internal progress, limited by the current methods and available resources, while external targets, defined by stakeholders, e.g., society or companies, drive technicians. This description matches the idea that the usefulness of an analysis should be evaluated within the context of its goals [11] and distinguishes two types of anomalies: "Scientific" anomalies of a novel observation and "technical" anomalies as deviations from a predefined norm using a predefined measurement of substantiality.

"Scientific" anomalies might still result in unwanted system states, which then can result in some kind of error or critical system state. Nevertheless, not every "scientific" anomaly inevitably results in an error state and not every error state is a "scientific" anomaly. An anomaly is not a "scientific" anomaly if the error state is already documented or can be described by the standard model. In this case, the anomaly becomes a "technical" anomaly.

Using the philosophy of science definition of anomalies, the normal state is the prediction by the system model, the deviation is the difference between the prediction of the system state and the measured actual state of the system, and the substantiality is defined by the noise and precision of our predictions and measurement tools.

3 Proposed Framework for a Formalization of Anomalies

To separate "scientific" and "technical" anomalies, a formerly proposed framework [5] is generalized as illustrated in Fig. 2. and mathematically defined in this section.

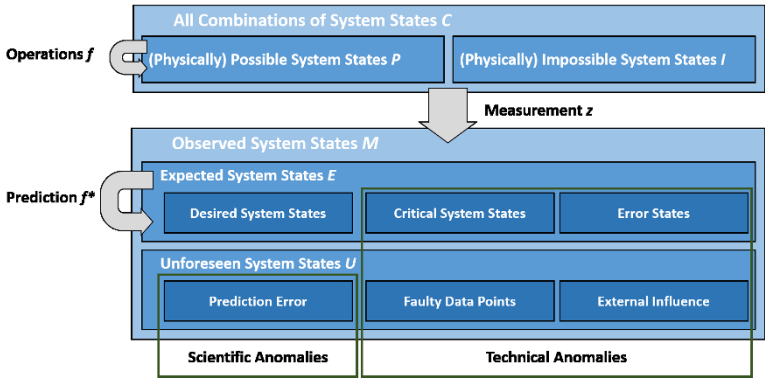

Fig. 1 Formalization of "scientific" and "technical" anomalies and system states.

Definition 1 (System State) There exists a multivariate description x_i of a state i with a finite number of features. For each feature j of state i a value x_{ij} exists, which is a realization of the feature space R_j. The value x_{ij} is the actual and precise state description of feature j at state i. Although there exists only a single true value x_{ij}, the value itself does not necessarily have to be a single data point but can be a multivariate or symbolic data value and can be of any data type.

$$\forall i \, \forall j \, \exists! \, x_{ij}, \quad x_{ij} \in R_j \tag{1}$$

The set C of all combinations of system state values with J features is given by:

$$C = \{x_i \mid \forall j \, \exists \, x_{ij} \in R_j\} = R_1 \times \ldots \times R_J \tag{2}$$

Definition 2 (Operation) An operation is an analytical function f which changes the system state from state i to the following state $i + 1$. Both states belong to the set of all combinations of system states C.

$$f : C \rightarrow C, \quad f(x_i) = x_{i+1} \tag{3}$$

There exists a finite set F of functions of endogenous state transformations. This set of functions is the scope of operations that can be performed. These functions are the fundamental functionality of a system, which can be performed without any external involvement. For all functions the following expression is applied:

$$g \in F \wedge f \in F : g \circ f \in F \tag{4}$$

Using the defined function space, a restriction of reachable system states via all functions from F is defined, resulting in the set of physically possible system states.

Definition 3 (Physically Possible System States) The relation f spans the complete space of state changes of a system using the entire scope of operations. The resulting space is the set of all possible system states. The physically possible system states

are the possible realizations of x_i based on a starting point and if only functions from F are applied. The set P is a group with a neutral element of operations.

$$P = \{x_i \mid \forall f \in F : f(x_i) \in P\} \subseteq C \tag{5}$$

Definition 4 (Observed System States) Of the amount J of existing features of the system state, only an amount D of features is known with $D \leq J$. Since not all system states can be measured, a function z transforms the real system states and real operations of the system into observable system states and operations.

$$z : C \rightarrow M, \quad z(x_i) = x_{i^*} \tag{6}$$

Therefore, the set $M = R_1 \times ... \times R_D$ is the space of all observable and known system states. Function z is the measurement process.

Definition 5 (Observed Operations) Not all functions of the whole set of function F are known or observable when planning and operating a system.

$$F' \subseteq F \tag{7}$$

Additionally, only observable system states are modeled when operating a system. The observed operations of systems are therefore projections of a subsets of known operations of F and operate within the observed and known system states.

$$F^* = z(F') \tag{8}$$

The actual conducted operations f are always from the set of operations F, but the expectation and prediction utilize, due to lack of system knowledge, only $f^* \in F^*$.

$$f^* : M \rightarrow M, \quad f^*(x_{i^*}) = x_{i+1^*} \tag{9}$$

Therefore, all states applied in operation f^* are defined as expected system states.

Definition 6 (Expected System States) The system states, which are possible if only the observed and known operations of the set F^* are applied to all system states $x_{i^*} \in E$, are the expected system behavior.

$$E = \{x_{i^*} \mid \forall f^* \in F^* : f^*(x_{i^*}) \in E\} \subseteq M \tag{10}$$

The expected system states can be further split into desired system states, where the system is running most beneficially for its usage, a critical system state, where a possible error or rare system states are measured, and error states, which are system faults with operational risks involved as defined by Basel III [18]. Applied in engineering, this definition is compatible with the definition of DIN EN 13306 since the system is at risk of being unable to perform a certain range of functions without necessarily being completely inoperable [19]. All kinds of errors, warnings and non-beneficial system states are the "technical" anomalies within the contextual analysis of the data set.

Definition 7 (Unforeseen System States) The set of unforeseen system states U are therefore all measurable system states within the realm of observable system states but not within the expected system states:

$$U = M/E \tag{11}$$

"Scientific" anomalies in unforeseen system states are measured if the real operation f differs from f^* such that a prediction error occurs:

$$f^*(x_{i^*}) \in E, \quad f^*(x_{i^*}) \neq z(f(x_i)) \notin E \tag{12}$$

"Scientific" anomalies are part of the unforeseen system states. Another reason for unforeseen system states is a measurement of an impossible system state. Anomalies originated by physically impossible system states are to be distinguished from "scientific" anomalies since the reason for their occurrence follows a different mechanism. Thus, they are assigned to the "technical" anomalies.

Definition 8 (Physically Impossible System States) Physically impossible system states I are combinations of states in set C which are not reachable using function f:

$$I = C/P \tag{13}$$

Definition 9 (External Influence) Applying changes to the system, the feature space also changes. Consequently, the space of the physically possible system states changes. Previously impossible system states become possible system states.

Definition 10 (Faulty Data Points) If a measurement is conducted incorrectly, the measured values could be within the impossible system states. Faulty data points are therefore neither measurement noise nor imprecision, but should be systematically excluded. Note that faulty data points could be within the possible system space but need to be excluded either way.

4 Conclusion

It is concluded that the anomaly concept is often loosely defined and heavily depends on assumptions of a normal state, deviation, and substantiality. These definitions are often case-specific and influenced by the conducting researchers' choice. Therefore, a rigorous definition of anomalies is capable of further streamlining the discourse and increasing a common understanding of what kind of anomaly is described.

Using "technical" and "scientific" anomalies, further research will be conducted to set up models detecting both types of anomalies separately. Differences between observed and real system states and operations are a focus of further research to more precisely analyze the hidden processes of the "scientific" anomaly generation. Also, a more fundamental discussion of the philosophical definition of anomalies within the philosophy of science and its applications to anomaly detection in general should be conducted to further gain insight into the true nature of anomalies.

The authors plan to validate the concept by using the proposed definition and framework in exemplary applications within industrial processes. Furthermore, anomaly detection methods designed for applications in dynamical systems using the proposed framework are planned to be developed.

Acknowledgements The Mercedes-Benz Group AG funds this research. The research was prepared within the framework of the doctoral program of the Institut für Informationsmanagement im Ingenieurwesen (IMI) at the Karlsruhe Institute of Technology (KIT).

References

1. Aggarwal, C. C.: Outlier Analysis. Springer Science+Business Media, New York (2013)
2. Hodge, V. J., Austin, J.A.: Survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85-126 (2004)
3. Aggarwal, C. C., Sathe, S.: Outlier Ensembles. Springer, Cham (2017)
4. Wang, X., Wang, X., Wilkes M.: New Developments in Unsupervised Outlier Detection - Algorithms and Applications. Springer, Singapore (2021)
5. Spoor, J. M., Weber, J., Ovtcharova, J.: A definition of anomalies, measurements and predictions in dynamical engineering systems for streamlined novelty detection. Accepted for the 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul (2022)
6. Åström, K. J., Murray, R. M.: Feedback Systems - An Introduction for Scientists and Engineers. Princeton University Press, Princeton, New Jersey (2008)
7. Sethi, S. P., Thompson, G. L.: Optimal Control Theory - Applications to Management Science and Economics. Springer Science+Business Media, Boston, MA (2000)
8. Mehrotra, K. G., Mohan, C., Huang, H.: Anomaly Detection - Principles and Algorithms. Springer International Publishing, Cham (2017)
9. Skiena, S. S.: The Data Science Design Manual. Springer International Publishing, Cham (2017)
10. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer Science+Business Media, New York (2013)
11. Fahrmeier, L., Hamerle, A., Tutz, G. (ed.): Multivariate Statistische Verfahren. de Gruyter, Berlin (1996)
12. Rousseeuw, P. J., Leroy, A. M.: Robust Regression and Outlier Detection. John Wiley & Sons, Inc (1987)
13. Foorthuis, R.: On the nature and types of anomalies: A review of deviations in data. Int. J. Data Sci. Anal. **12**, 297-331 (2021)
14. Cuadrado-Gallego, J. J., Demchenko, Y.: The Data Science Framework: A View from the EDISON Project. Springer Nature Switzerland AG, Cham (2020)
15. Kuhn, T.: The Structure of Scientific Revolutions. 2nd ed. The University of Chicago Press, Chicago (1970)
16. Hawkins, D.: Identification of Outliers. Chapman and Hall (1980)
17. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) 15 (2009)
18. Bank for International Settlements: Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards (2006)
19. DIN Deutsches Institut für Normung e. V.: DIN EN 13306: Instandhaltung - Begriffe der Instandhaltung. Beuth Verlag GmbH, Berlin (2010)