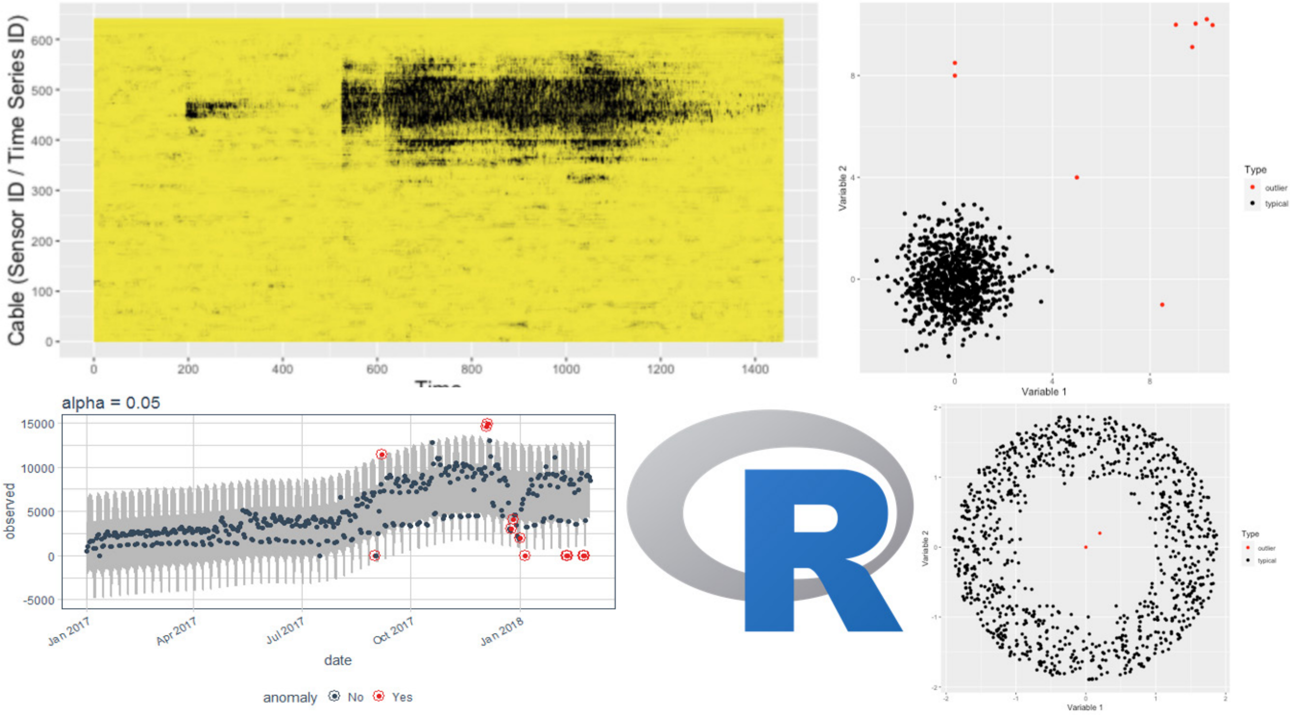
Unveiling the Unusual: A Task View for Anomaly Dection in R



Anomalies play a critical role in statistical analysis, as their presence in data can lead to biased parameter estimation, model misspecification, and misleading results if classical analysis techniques are blindly applied. Additionally, anomalies can themselves be carriers of significant and critical information, and identifying these critical points can be the primary goal of investigations in many fields such as fraud detection, object tracking, system health monitoring, and environmental monitoring (e.g., for bushfires, tsunamis, floods, earthquakes, and volcanic eruptions) (10).

Anomaly detection problems have multiple facets, and detection techniques can be highly influenced by the way anomalies are defined, the type of input data used in the algorithm, the expected output, and more. These factors lead to wide variations in problem formulations that require different analytical approaches. However, many existing analytical approaches are out of reach for applied researchers and mostly limited to their original applications due to their complexity, difficulty, and the time required to implement them in standard statistical software distributions. With the exponential growth of data and associated research challenges, it is more efficient and effective to use statistical software that facilitates anomaly detection capabilities. Commercial and freely available software packages have been developed to provide anomaly detection methods with varying degrees of capabilities, functionality, and ease of use.

R is a programming language and open-source software managed and maintained by the Comprehensive R network (8). Among the many software possibilities available, R has become an increasingly popular choice of software in both academic and applied settings. It facilitates an environment for academics, data scientists, and statisticians to provide state-of-the-art implementations of newly developed techniques, and it helps applied researchers handle their research questions through various built-in functions. Another factor contributing to R's success is its community, which consists of leading scientists who have developed good and reliable software packages (9).

The lack of a unifed defnition for an anomaly has contributed to the introduction of many different R packages with anomaly detection capabilities. At present, there is a wide variety of R software packages available, exceeding 100 in number, that support anomaly detection tasks across different disciplinary contexts utilizing a range of analytical techniques. Despite this, these packages generally offer an easy-to-use environment for anomaly detection with minimal computational effort. The development of this growing number of R packages for anomaly detection is the result of efforts from different independent contributors motivated by various applications in different fields. As a consequence of this development dynamics and the diversity of contributed packages, the available functionality in R packages is not uniform in usage, although some do share similarities in their broader framework. This diversity can make it challenging for users to determine which package to use for their specific problem. Furthermore, as R allows users to access these packages from different sources, such as CRAN, GitHub, and Bioconductor (BioC), keeping track of this evolution can be challenging for both users and developers.

Choosing the right R package for anomaly detection can be a daunting task, requiring careful consideration of project requirements and package capabilities. However, with a large and constantly evolving collection of packages available, it can be difficult for users to know which packages to choose, and how to best make use of their features and limitations (4). To help users navigate this landscape, we have introduced a CRAN task view dedicated to anomaly detection in R. This task view provides a comprehensive, up-to-date catalog of publicly available R packages for anomaly detection, along with detailed information on the kinds of support and capabilities that each package offers. The task view is available for review and use on GitHub, at https://github.com/pridiltal/ctv-AnomalyDetection. The dynamic nature of this task view allows researchers to keep themselves updated with the latest developments and releases related to anomaly detection.

To identify and collect all publicly available R packages for anomaly detection, we regularly perform a systematic and comprehensive search. Anomalies are also referred to as outliers, novelty, faults, deviants, dis- cordant observations, extreme values/cases, change points, events, intrusions, misuses, exceptions, aberrations, surprises, peculiarities, odd values, or contaminants in various application domains (1,4). Therefore, we began our search by exploring three standard websites that host R packages: the Comprehensive R Archive Network (CRAN) website, Crantastic (11), Bioconductor (3), and GitHub (github.com), using these keywords and their variations. In addition to these sites, we also gathered information from other sources, such as the package update feed on CRAN, Google Scholar (scholar.google.com), blog posts that provide links and information about anomaly detection methods and related applications, and email communications from relevant user groups and developers of available R software packages for anomaly detection. This allowed us to compile a diverse collection of information on publicly available R packages for anomaly detection. We also update the task view on monthly basis to reflect any changes in the state of the field, new techniques, or emerging trends. This helps ensure that the task view remains a valuable resource for users looking to stay up-to-date with the latest developments in their field.

We also found packages in R that allow for conducting anomaly detection tasks as secondary operations. For instance, the forecast R package (5) is primarily a suite of functions for univariate time series forecasting. However, the package also includes a function called tsoutliers for identifying outliers in time series data and proposing reasonable replacements. If a function has a theoretical basis for the anomaly detection task, we include it in the task view, regardless of the package's primary focus. Conversely, if an R package for anomaly detection provides different functionality that deviates from our main focus, we report this feature but do not review it since our main focus is solely on anomaly detection.

The formulation of an anomaly detection problem and the possible solutions depend on a variety of factors, including the application domain and the constraints and requirements associated with it (2). The structure of our CRAN task view reflects this complexity and includes 2 sections on different facets of the topic, such as the nature of the input data; modeling approaches used for different data structures (univariate, multivariate, temporal, spatial, spatio-temporal, and functional); methods for calculating anomalous thresholds; and target applications. Worth noting that these different facets are not mutually exclusive, and some packages may appear in multiple sections of the task view. To ensure that the CRAN task view remains up-to-date and relevant, we periodically updates the list of packages with any new or updated packages that are relevant to the topic area. This process is ongoing and involves monitoring new package submissions, reviewing existing packages, soliciting feedback, and publishing updates.

We believe that our CRAN Task Views serve as a valuable resource for researchers seeking to stay up-to-date with the latest developments in their field and to learn about new R packages that are relevant to their work. Our review of anomaly detection packages in R aims to simplify the process of selecting the most suitable package for a user's specific anomaly detection needs. Furthermore, we hope that our task view will provide insight into the various directions in which research has been conducted in this area, enabling users to combine the functionality of multiple packages to achieve desired outcomes.

During our survey we noticed a significant amount of overlap in the functionality provided by the R packages for anomaly detection, and we believe that our regularly updated task view will encourage future package developers to focus on improving existing capabilities rather than duplicating efforts that have already been addressed. Additionally, we recognized that there are still anomaly detection tasks and methodological approaches that are not adequately supported by the current collection of R packages, and we hope that our CRAN task view for anomaly detection will inspire researchers and package developers to explore more advanced methods and bridge existing gaps in anomaly detection applications within the R environment.

This work was carried out with the aid of a grant from UNESCO and the International Development Research Centre, Ottawa, Canada. The views expressed herein do not necessarily represent those of UNESCO, IDRC or its Board of Governors.

References:

[1] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. "Anomaly Detection: A Survey." ACM Computing Surveys 41 (3): 1–58.

[2] Filzmoser, Peter, Robert G Garrett, and Clemens Reimann. 2005. "Multivariate Outlier Detection in Exploration Geochemistry." Computers & Geosciences 31 (5): 579–87.

[3] Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." Genome Biology 5 (10): R80.

[4] Gupta, Manish, Jing Gao, Charu Aggarwal, and Jiawei Han. 2014. "Outlier Detection for Temporal Data." Synthesis Lectures on Data Mining and Knowledge Discovery 5 (1): 1–129.

[5] Hyndman, Rob J, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava

[6] Razbash, Earo Wang, and Farah Yasmeen. 2018. "Forecast: Forecasting Functions for Time Series and Linear Models." http://pkg.robjhyndman.com/forecast.

[7] Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. 2010. "Outlier Detection Techniques." Tutorial at KDD 10.

[8] R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

[9] Rusch, Thomas, Patrick Mair, and Reinhold Hatzinger. 2013. "Psychometrics with r: A Review of CRAN Packages for Item Response Theory."

[10] Talagala, Priyanga Dilini, Rob J Hyndman, and Kate Smith-Miles. 2021. "Anomaly Detection in High- Dimensional Data." Journal of Computational and Graphical Statistics 30 (2): 360–74.

[11] Wickham, H., and B. Maeland. 2009. "Crantatistic." https://www.crantastic.org/.

Article by

Priyanga Talagala

Department of Computational Mathematics, Faculty of Information Technology, University of Moratuwa, Sri Lanka