Why is Greenwich so common? Quantifying the uniqueness of multivariate observations

Andrea Ballatore
Department of Digital Humanities, King's College London, UK

Stefano Cavazzi
Ordnance Survey, UK

Abstract

The concept of uniqueness can play an important role when the assessment of an observation's distinctiveness is essential. This article introduces a distance-based uniqueness measure that quantifies the relative rarity or commonness of a multi-variate observation within a dataset. Unique observations exhibit rare *combinations* of values, and not necessarily extreme values. Taking a cognitive psychological perspective, our measure defines uniqueness as the sum of distances between a target observation and all other observations. After presenting the measure u and its corresponding standardised version u_z, we propose a method to calculate a p value through a probability density function. We then demonstrate the measure's behaviour in a case study on the uniqueness of Greater London boroughs, based on real-world socioeconomic variables. This initial investigation indicates that u can support exploratory data analysis.

2012 ACM Subject Classification Mathematics of computing → Multivariate statistics; Information systems → Geographic information systems

Keywords and phrases uniqueness; distinctiveness; similarity; outlier detection; multivariate data

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.21

Supplementary Material All R code and data used to produce this article are available on GitHub at `https://github.com/andrea-ballatore/calculating-uniqueness`.

Acknowledgements This work was supported by Ordnance Survey.

1 Introduction

The identification of similar observations is a well-studied problem in data science [2]. All forms of clustering rely on some form of similarity assessment [7], and so does data deduplication. Online platforms relentlessly search for similar products and similar users to drive engagement and sales. Geographic concepts can be compared and grouped by similarity and relatedness [1]. Geo-demographic classifications group similar areas based on socioeconomic characteristics. Similarity enables the identification of points representing points that appear close in a multi-dimensional vector space [7].

But what about uniqueness, one of the similarity's less known siblings? In many domains, the degree to which an object is unique is crucial to assess its value. The distinctiveness of artworks is carefully studied by scholars to ascertain the originality of painters, musicians, and writers. Unique cities, landscapes, and heritage assets are praised in the rhetoric of tourism marketing [9]. In the natural sciences, uniqueness is useful to define physical or chemical properties, genetic or molecular characteristics, or ecological traits that distinguish an individual from all others. Unique fingerprints, faces, irises, and DNA sequences enable ubiquitous applications in cybersecurity and forensic science. The cognate concept of "distinctiveness" is used in biology to explore the taxonomic structure of species [4]. In recommender systems, it has been deployed to assess the typicality of user preferences [8]. The uniqueness of observations has been occasionally operationalised to support the interpretation

and filtering of multi-dimensional data [5]. In its infrequent appearances in the scientific literature, this concept is largely left unexamined.

A relevant and intensely investigated problem is that of outlier detection, which consists of identifying unusual observations that might indicate infrequent but interesting events (e.g., fraudulent bank transactions) or measurement errors. Outliers can be found with distance-, clustering-, density-, ensemble-, and learning-based methods, with varying levels of success and robustness [10]. In a multivariate context, outliers are *rare combinations of values*. The values of each variable might not be extreme, but the combination appears relatively far from the others. The Mahalanobis distance is very useful for finding outliers in multidimensional datasets. While linked to outlier detection, our objective is the quantification of uniqueness as a facet of observations for classification and exploratory data analysis.

To support the exploration and recommendation of walking routes [3], we devised a distance-based uniqueness measure that can quantify how relatively rare (or common) an observation is in a dataset. In the spirit of classic ecological indices that have been in vogue for 40 years [11], we devise a simple, general, and easily interpretable measure that can be applied to many contexts. We ground our notion of uniqueness into a cognitive psychological perspective that defines the "distinctiveness of stimuli" as "the sum of the differences between the stimulus and all other stimuli in the group" [6, p. 16]. Hence, the more a multi-variate observation is different to all others, the more it is unique. The relative rarity of observations can act as an informative feature in machine learning methods and can support user interaction and data interpretation. Our measure u is described in the remainder of this article. Its behaviour is illustrated in a case study on London boroughs.

2 A uniqueness measure

Univariate uniqueness

In its simplest form, uniqueness can be thought of as $1 - p$, where p is the probability of encountering a particular observation from random extractions from a set. For example, let us consider the percentages of land cover categories of the UK territory: *farmland* 56.7%, *natural* 34.9%, *green urban* 2.5%, *built* 5.9%.[1] Taking this probabilistic view, the rarest category we would encounter by selecting a random area is green urban, corresponding to $u = .98$, and the most common (least unique) category is farmland ($u = .43$). This is conceptually linked to the idea of surprise—a less likely outcome is more surprising.

To make u more interpretable, we can calculate the corresponding z scores, relating uniqueness to the deviation from the normal distribution—with an assumption of normality that might not hold. If all types of observation occur at the same probability, it is not possible to meaningfully calculate uniqueness (z is null). Otherwise, common observations have negative z scores, and rare ones are positive: *farmland* has $z = -1.24$, while *green urban* has the highest value, with $z = .88$.

Multivariate uniqueness

As part of our efforts to support the exploration of large datasets [3], we developed a uniqueness measure that can handle multivariate observations. Considering a set of observations, the frequency of a given multi-variate configuration is correlated to uniqueness as rare observations are more unique than common ones. From a statistical standpoint, the assessment of

[1] `https://www.eea.europa.eu/publications/CORO-landcover`

uniqueness is also analogous with outlier detection, in which observations at the extreme of a distribution can be identified as of particular interest or as the result of measurement errors.

Our uniqueness index u between a multi-variate observation is calculated as the sum of the similarity of the observation with all other observations in the group S, ranging from rare to very common. Formally, given a set of observations S, the uniqueness score u of an observation $a \in S$ is defined as:

$$u(a,S) = \sum_{i=1}^{|S|} d(a,a_i), \quad a \neq a_i,\ d \geq 0,\ u \geq 0,\ S = \{a_1 \cdots a_m\}, \quad u_z = \frac{u - \hat{u}}{\sigma(u)}$$

where d is an n-dimensional distance function. Different functions, such as Euclidean, Manhattan, or Mahalanobis, might produce radically different u. This measure is also sensitive to the particular structure of the data and to the selected variables. The scores are then standardised as u_z as z scores, where $\hat{u}$ is the mean u and $\sigma(u)$ the standard deviation. The u_z scores are more interpretable than u, as they embed a measure of distance from the dominant clusters in the data. In other words, the index allows comparing observations on a spectrum ranging from very common (low values) to very rare (high values of u_z). An intuitive interpretation of these scores relates to the distance from cluster centres in the data space: Central data is common, and peripheral is rare.

In order to provide a measure of statistical significance, we calculate a p value for each standardised $z(u)$ using a probability density function of a normal distribution, defined as $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$, with $\hat{u_z}$ as the mean μ and σ_{u_z} as σ. This approach captures the extent to which a value of u_z is greater than expected, with lower p for rarer cases. The underlying assumption is that u scores have an approximately normal distribution. This is a simplification of real-world data that might exhibit very different distributions of u and should be adjusted to specific contexts, but it is useful to develop our measure.

The behaviour of u_z and p values was tested on synthetic datasets of multivariate data, with a Monte Carlo approach, generating random high-dimensional datasets with different distributions and calculating u_z and corresponding p values, considering uniform, normal, and clustered distributions with two and three large clusters. In this initial empirical investigation, the observation-by-attribute matrices showed that the distribution of uniqueness scores remains fairly stable across different matrix sizes and across different distributions, although low p values are more frequently produced with clustered data. For example, considering 1440 matrices, on average, a normal distribution produced 0.54% of p smaller than .001 and 89% at $p > .1$, which makes intuitive sense.

Interpreting uniqueness

In a focus group we conducted at Ordnance Survey [3], we discussed the semantic interpretation of these scores with stakeholders. The uniqueness measure u_z was presented with walking routes as items to score, based on a number of attributes to identify unusual routes. The term "uniqueness" was considered semantically clearer than "distinctiveness." From a cognitive perspective, participants expressed a preference for a categorical classification as opposed to both scores and ranks. Less agreement was found on the specific categories to use. The terms discussed included "common", "rare", and "typical" with modifiers "very" and "extremely". It was noted that terms should not have positive or negative connotations, devaluing common items as uninteresting or valuing rare items that might be uncommon for good reasons—an unusual walking route around a landfill. Moreover, the participants highlighted the importance of showing the discriminant attributes along with the scores.

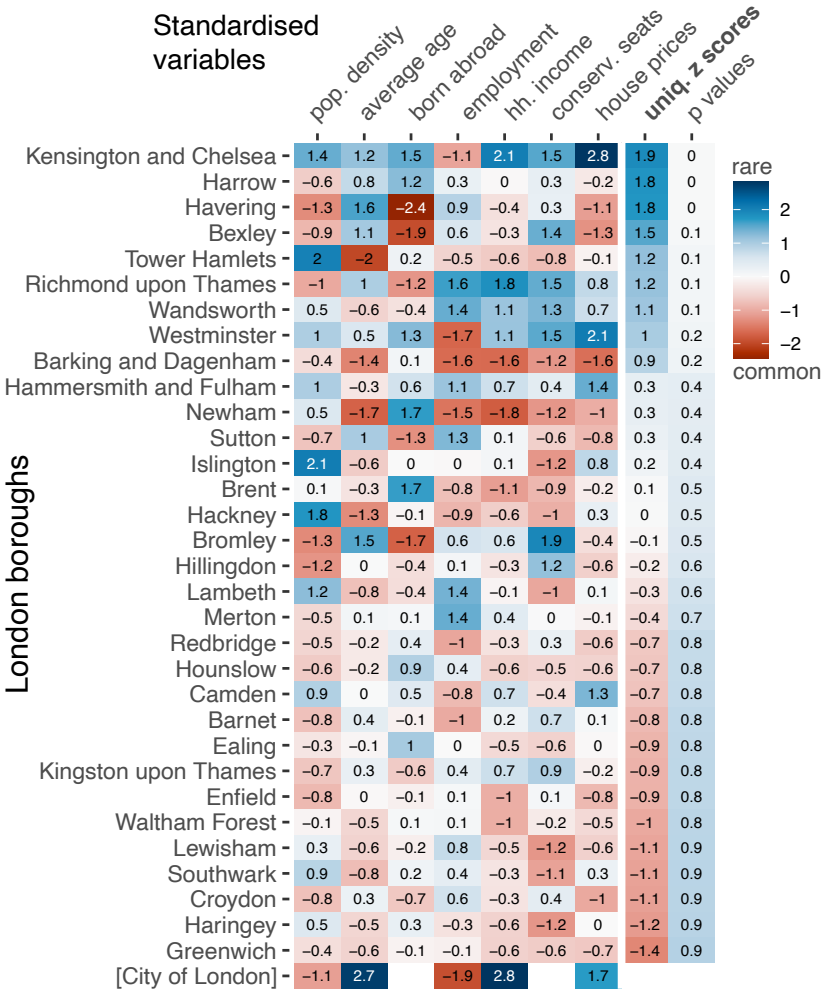

Figure 1 Uniqueness of 32 London boroughs with respect to seven socioeconomic variables, excluding City of London, calculated with the Mahalanobis distance. Variables were centred and standardised. The rows are ordered by the uniqueness z scores (u_z) from the rarest (Kensington and Chelsea) to the most common (Greenwich). Data source: London borough profiles, 2015.

As a result of this process, we defined five uniqueness levels based on p values as follows: ($p = 0$) *very rare* (.001) *rare* (.01) *intermediate* (.05) *common* (.1) *very common* (1). For example, $p = .006$ would be classified as rare. While such classifications are inevitably domain-dependent, these bins appear easily interpretable as they segment the scores at common p value thresholds.

3 The uniqueness of London boroughs

As an exploratory case study, we consider the boroughs of Greater London, a familiar, well-understood geography described through a set of socioeconomic variables. The seven selected variables include population density, average age, percentage of residents born abroad, percentage of employed residents, household income, Conservative seats, and median

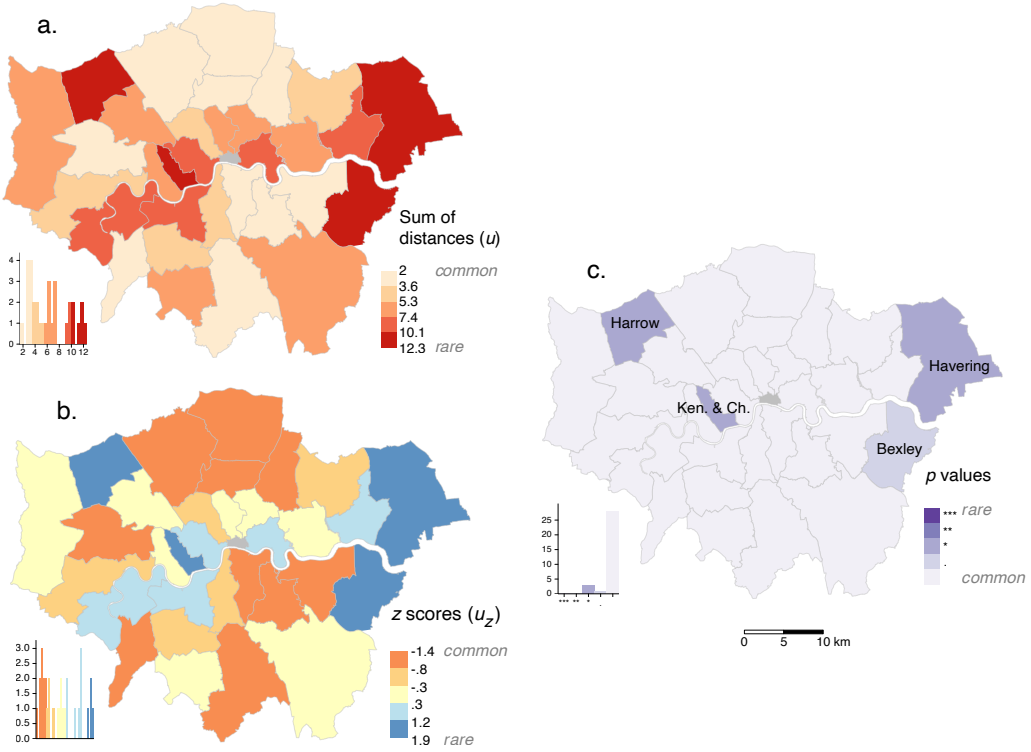

Figure 2 Uniqueness of 32 London boroughs with respect to seven socioeconomic variables, excluding City of London, calculated with Mahalanobis distance. (a) Sum of distances u; (b) Uniqueness z scores (u_z); (c) p values with thresholds . ($p < .1$), * ($p < .05$), ** ($p < .01$), *** ($p < .001$). Bins for (a) and (b) were produced with Jenks. Data source: London borough profiles, 2015. Projection: British National Grid (EPSG:27700).

house price.[2] Highly correlated variables such as Labour seats ($\rho < -.7$ or $\rho > .7$) were removed to avoid obvious collinearity issues. Using our R implementation, we performed calculations of u, u_z, and p values for all boroughs, except for the City of London, which had several missing values. Among many possible options, the Mahalanobis distance measure was selected as it inherently accounts for scale invariance. Figure 1 presents the standardised matrix used for the calculation, along with the corresponding u_z and p values.

Boroughs exhibiting the highest u values are characterized by unexpected combinations of variables. According to our calculation method, three boroughs stand out as significantly rare, with uniqueness scores ($p < .05$). Kensington and Chelsea demonstrate exceptionally high house prices and household income, but relatively low levels of employment. Harrow, on the other hand, is intriguing as its variables are not extreme individually, but noticeably distinct from all other boroughs as a whole. Lastly, Havering appears relatively unique due to its ageing population and predominantly UK-born residents. In contrast, Croydon, Haringey, and Greenwich, located towards the bottom of the matrix, exhibit more central positions in the data, making them more representative of London as a whole. Greenwich, at least based on these variables, emerges as a very typical—and therefore common in our parlance—borough of Greater London. Figure 2 displays maps illustrating the spatial

[2] Data source: London Borough Profiles and Atlas, Greater London Authority (GLA), 2015.

distribution of u, u_z, and p values. Visually, the three rare boroughs do not exhibit clustering.

The results from this analysis indicate that our measure, u, shows promise in quantifying the uniqueness of multivariate observations. However, further empirical testing with both real-world and synthetic data is necessary to assess the stability and interpretability of this uniqueness measure across different domains. A noteworthy characteristic of u is its weak correlation with any of the seven variables (the strongest correlation being $\rho = .39$). This suggests that the measure is capturing a latent dimension of the data, i.e. the distribution of uniqueness of these observations, revealing this facet of the data for further analysis.

In conclusion, further empirical testing is necessary to evaluate the stability and cognitive plausibility of this uniqueness measure across domains. Comparing different distance measures and methods for calculating p values is crucial to assess u's sensitivity to minor data variations. The operationalisation of uniqueness might support meaningful analyses of why some places, cultural artefacts, human behaviours, and natural environments emerge as unique from a vast sea of sameness.

References

1. Andrea Ballatore, Michela Bertolotto, and David C. Wilson. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, 18:747–767, 2014. doi:10.1007/s10707-013-0197-8.
2. Andrea Ballatore, Michela Bertolotto, and David C Wilson. A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS International Journal of Geo-Information*, 4(2):471–492, 2015. doi:10.3390/ijgi4020471.
3. Andrea Ballatore, Stefano Cavazzi, and Jeremy Morley. The context of outdoor walking: A classification of user-generated routes. *The Geographical Journal*, 2023. doi:10.1111/geoj.12511.
4. K Robert Clarke and Richard M Warwick. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35(4):523–531, 1998.
5. Pamela J Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 631–638, New York, 2004. ACM.
6. Bennet B Murdock Jr. The distinctiveness of stimuli. *Psychological review*, 67(1):16–31, 1960.
7. Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*, 2022.
8. Haggai Roitman, David Carmel, Yosi Mass, and Iris Eiron. Modeling the uniqueness of the user preferences for recommendation systems. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 777–780, New York, 2013. ACM.
9. Jonathan Schifferes. Mapping heritage. *RSA Journal*, 161(5563):10–13, 2015.
10. Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000, 2019.
11. HG Washington. Diversity, biotic and similarity indices: A review with special relevance to aquatic ecosystems. *Water Research*, 18(6):653–694, 1984.