Journal of Classification
https://doi.org/10.1007/s00357-025-09508-3

ORIGINAL RESEARCH

On the Local Ultrametricity of Finite Metric Data

Patrick Erik Bradley[1]

Accepted: 13 March 2025

Abstract
New local ultrametricity measures for finite metric data are proposed through the viewpoint that their Vietoris-Rips corners are samples from p-adic Mumford curves endowed with a Radon measure coming from a regular differential 1-form. This is experimentally applied to three datasets.

Keywords Local ultrametricity · p-adic numbers · Finite data · Mumford curves · Vietoris-Rips complex · Data analysis

1 Introduction

Ultrametricity is appealing for many reasons, and in particular, the simplicity of tree structures encoded in ultrametric spaces seems attractive to data analysts. Because of this, they would like to see how close to an ultrametric space a given data set is in order to extract something meaningful out of a hierarchical classification of the data. With this in mind, ultrametricity indices have been proposed, e.g., by Rammal et al. (1986) or Murtagh (2004). F. Murtagh observed experimentally that samples which are sparse and random in hypercubes become more and more ultrametric as dimension increases, using his ultrametricity index (Murtagh, 2004). Explanations for this are given in Bradley (2016) and Zubarev (2014). Also, ultrametricity can be related to topological data analysis (Bradley, 2017), and a corresponding ultrametricity index has a logistic behaviour (Bradley, 2019). This index relies on the Vietoris-Rips complex developed in Vietoris (1927), which is important in studying the persistent homology of data. Cf., e.g., Zomorodian (2010) for a fast construction of the Vietoris-Rips complex.

The p-adic numbers, having an inherent regular hierarchical structure, provide a framework for analysing hierarchical data, and thus, p-adic encoding methods were devised, Murtagh (2016) or Bradley (2010), either in order to bring them closer to ultrametricity or to apply p-adic methods to their already existing hierarchical structure. This leads to the applicability of p-adic analysis outlined, e.g., in Vladimirov et al. (1994) to the investigation

✉ Patrick Erik Bradley
bradley@kit.edu

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Englerstr. 6, 76131 Karlsruhe, Germany

of data. Applications outside of physics of p-adic analysis can be found in theoretical biology (cf. the review article (Dragovich et al., 2021), video data recognition (Benois-Pineau & Khrennikov, 2010), or medical diagnostics (Shor et al., 2021)). What is common to those applications is that they encode a hierarchical classification of a dataset. Since a hierarchical classification aims to find a suitable ultrametric for given data, the result is an embedding of hierarchically classified data into the p-adic numbers via a p-adic encoding. The main problem with a hierarchical classification of a whole dataset is that in reality, the actual similarities within data might be only of a local nature. That is why in this article a locally hierarchical classification is advocated. This also leads to p-adic encodings, like in the globally hierarchical case, but here, it leads to ones in which a hierarchical organisation of the "clusters", i.e., the maximal hierarchical pieces, within the space of p-adic numbers, is ignored. For this reason, the term local ultrametricity is used in this article.

The scope of this article is to introduce new measures for local ultrametricity, arguing that the clusters appearing as connected components of the Vietoris-Rips graphs from topological data analysis are likely to be more ultrametric than the whole dataset, which can be seen in an example case taken from the well-known iris dataset. Whether or not this argument is generally valid or not, the viewpoint induced by this approach leads to the idea that data can be seen as being sampled from Mumford curves. These are p-adic compact algebraic manifolds of dimension 1. Locally, they are holed discs in the p-adic number field, on which there is a natural Haar measure. However, the irregular tree structure of the local data leads to a more natural Radon measure coming from an algebraic regular differential 1-form on the Mumford curve, as constructed in Bradley and Ledezma (2024). There, the subdominant ultrametric associated with a finite metric space is used, which can be calculated with the method of Rammal et al. (1986). In fact, any ultrametric can be used to approximate the finite metric dataset, i.e., any hierarchical classification method can be used in order to obtain an ultrametric in this approach.

Mumford curves are objects studied in p-adic algebraic and rigid geometry and are extensively covered in Gerritzen and van der Put (1980) and Fresnel and van der Put (2004). What is needed from this relatively deep theory is, however, only the fact that they are algebraic and have an underlying 1-dimensional compact p-adic manifold structure which allows for regular differential 1-forms ω, which are in fact algebraic. Locally, they are of the form

$$\omega(x) = f(x)\,dx$$

with an analytic p-adic-valued function f defined on the local piece U, and that these give rise to Radon measures on the Mumford curve outside the zeros of ω.

The following Section 2 defines local ultrametrics as well as local p-adic data encoding via tree embeddings and uses the Vietoris-Rips graph from topological data analysis in order to define new data invariants and associated Mumford curves endowed with a Radon measure. This is followed by Section 3 consisting of experiments. A conclusion is given in Section 4.

2 Finite Locally Ultrametric Spaces

After defining local ultrametrics in the following subsection, and local p-adic encodings of data via tree embeddings, new invariants of a finite metric space are defined via the Vietoris-Rips graphs and associating Mumford curves and Radon measures to local pieces in the last subsection of this section.

2.1 Local Ultrametrics

Let X be a finite set with a metric d on it. Fix $\epsilon > 0$, and let Γ_ϵ be the associated Vietoris-Rips graph with vertex set X. Let

$$d_\epsilon : X \times X \to \mathbb{R}_{\geq 0}$$

be the partial function which on each connected component C of Γ_ϵ is an ultrametric dominated by d. One can use for d_ϵ, e.g., the subdominant corresponding to the distance on X restricted to $C \times C$. But any other ultrametric dominated by d can also be used. Certain hierarchical clustering methods provide such an ultrametric, among which single-linkage clustering yields the subdominant ultrametric.

Let $\mathcal{C}(\Gamma_\epsilon)$ be the set of connected components of Γ_ϵ. Define a distance d'_ϵ on $\mathcal{C}(\Gamma_\epsilon)$ as

$$d'_\epsilon(C, C') = \min\left\{\epsilon' \mid \epsilon' \geq \epsilon : \exists \text{ an edge in } \Gamma_{\epsilon'} \text{ connecting } C \text{ and } C'\right\},$$

whenever $C \neq C'$. Then, define the function

$$\delta_\epsilon : X \times X \to \mathbb{R}_{\geq 0},\ (x, y) \mapsto \begin{cases} d_\epsilon(x, y), & \exists\, C \in \mathcal{C}(\Gamma_\epsilon) : x, y \in C \\ d'_\epsilon(C(x), C(y)), & C(x) \neq C(y), \end{cases}$$

where $C(z) \in \mathcal{C}(\Gamma_\epsilon)$ is the connected component containing $z \in X$. Clearly, δ_ϵ is a distance on X.

Definition 2.1 The distance δ_ϵ is called a local ultrametric on X. The pair (X, δ_ϵ) is called a locally ultrametric space.

A criterion for ultrametricity in terms of the Vietoris-Rips graphs is given in Bradley (2016, Lem. 2.2): the connected components of the Vietoris-Rips graphs are always cliques iff dataset is ultrametric. The subdominant ultrametric can also be described in terms of the Vietoris-Rips graphs (cf. Bradley (2016, Prop. 5.2)).

Example 2.2 Figure 1 (left) shows a graph with five vertices as a metric space. The Vietoris-Rips graph for $\epsilon = 1$ has four connected components (called clusters), and the inter-cluster graph obtained by identifying clusters with vertices of a new graph is depicted in Fig. 1 (middle). The corresponding local ultrametric δ_1 can be read off the two figures as follows: The only non-singleton cluster is $\{a, b\}$ where the distance between a and b is read off the

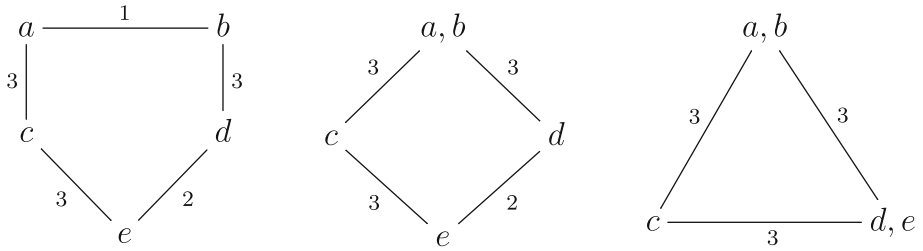

Fig. 1 Left: A graph as a metric space. Middle: The Vietoris-Rips inter-cluster graph for $\epsilon = 1$ (named Γ_1^3 according to Section 2.3). Right: The Vietoris-Rips inter-cluster graph for $\epsilon = 2$ (named Γ_2^3 according to Section 2.3)

left graph as 1. The other inter-cluster distances can be read off the graph in the middle by taking the length of the shortest path between clusters. The same holds true in the case $\epsilon = 2$, where now there are four clusters, two of which are non-singletons. Notice that the first Betti number of the two inter-cluster graphs is equal to one. This will be used to illustrate invariants in Section 2.3 below.

2.2 Local *p*-adic Encodings

In Bradley and Ledezma (2024, §3.3) a Radon measure on a compact open subset of $\mathbb{Q}_p$ is constructed from a finite ultrametric space. Here, an embedding of the corresponding ultrametric tree into the Bruhat-Tits tree of a suitable p-adic number field necessary for that method is constructed in a more precise manner. This produces a p-adic data encoding, as already observed in Bradley (2010).

Let $C \in \mathcal{C}(\Gamma_\epsilon)$ be given, and view (C, d_ϵ) as an independent ultrametric space for the moment. The set $\mathcal{B}(C)$ of all non-trivial balls on C is a finite poset with precisely one top element C and in fact is a tree. Let

$$\rho\colon \mathcal{B}(C) \to \mathbb{R}_{>0},\ B \mapsto \text{radius of } B\,,$$

whose image $R(C) = \rho(\mathcal{B}(C))$ is a finite ordered set of real numbers. Order this set with a function

$$\varphi\colon R(X) \to \mathbb{N}$$

in decreasing order with consecutive natural numbers beginning in 0. Fix a prime number p, and let

$$m = \max \varphi\,,$$

and assign to each $c \in C$ a distinct disc $a_c + p^{(m+1)}\mathbb{Z}_p$ inside the ring $\mathbb{Z}_p$ of p-adic integers inside the field of p-adic numbers $\mathbb{Q}_p$, where p is bounded from below by the maximal number of children in any ultrametric tree of (C, δ_ϵ) for any $C \in \mathcal{C}(\Gamma_\epsilon)$ plus the number of elements in $\mathcal{C}(\Gamma_\epsilon)$. Assume thereby that all discs in $\mathbb{Z}_p$ have equal radius

$$p^{-(m+1)}$$

for this assignment. The condition about p enables an embedding of any spanning tree of the graph Γ_ϵ into the Bruhat-Tits tree for $\mathbb{Q}_p$. The latter tree is explained, e.g., in Bradley (2006).

The ultrametric diffusion considered in Bradley and Ledezma (2024) necessitated the replacement of the p-adic Haar measure on the compact open set obtained by such an embedding as above with a Radon measure ν_C which ensures that the volume of any disc corresponding to a vertex $B \in \mathcal{B}(C)$ in the ultrametric tree for (c, d_ϵ) is equally distributed among the child vertices of B, cf. Bradley and Ledezma (2024, Lem. 3.8), where such a Radon measure is constructed on the subset

$$\Omega_C = \bigsqcup_{c \in C} \left(a_c + p^{m+1}\mathbb{Z}_p\right)$$

of $\mathbb{Q}_p$.

Definition 2.3 The measure ν_C is called the equity measure on Ω_C induced by (C, d_ϵ).

Now, Bradley and Ledezma (2024, Lem. 3.9) show that ν_C is of the form

$$\nu(x) = \phi(|f_C(x)|)\ |dx|_p\ ,$$

where $|dx|_p$ is the Haar measure on $\mathbb{Q}_p$, $f_C \in \mathbb{Q}_p[X]$ a polynomial nowhere vanishing on Ω_C, and $\phi\colon p^{\mathbb{Z}} \to \mathbb{R}_{>0}$ a strictly increasing function. The proof of Bradley and Ledezma (2024, Lem. 3.9) uses p-adic polynomial interpolation.

Example 2.4 Continuing Example 2.2, Fig. 2 (left) shows a global 2-adic encoding of the five vertices of the graph in Fig. 1 (left) which locally induces the subdominant ultrametric on the Vietoris-Rips clusters for $\epsilon = 2$. More precisely, a binary number encoding a vertex v is given by the geodesic path from root to v, and then taking

$$c(v) = \sum_{i=0}^{3} \alpha_i 2^i\ ,$$

where α_i is the number attached to the i-th edge downwards from the root vertex along this path. The local ultrametric distance within each cluster coincides with

$$d_1(x, y) = 2^3 \cdot \log|x - y|_2$$

in this example. The normalised Haar measure gives each vertex equal volume. In case the 2-adic unit disc is assumed to have volume 1, each vertex corresponds to a 2-adic disc of volume $\frac{1}{8}$. An equity probability measure on the dataset is given by assigning equal volume to each branch attached to a vertex in the ultrametric tree associated with each cluster. Since there are many possible choices for this, as it is only locally defined, this has been for the global dendrogram of all vertices given by the 2-adic encoding of Fig 2 (right) corresponding to the g ultrametric.

2.3 Invariants of a Finite Metric Space

Let $\delta \geq \epsilon$. Define the graph Γ_ϵ^δ whose vertex set is $\mathcal{C}(\Gamma_\epsilon)$, and its edges are pairs (C, C') with $C \neq C'$ and $d(C, C') \leq \delta$. We call this the coarse ϵ-δ graph of (X, d).

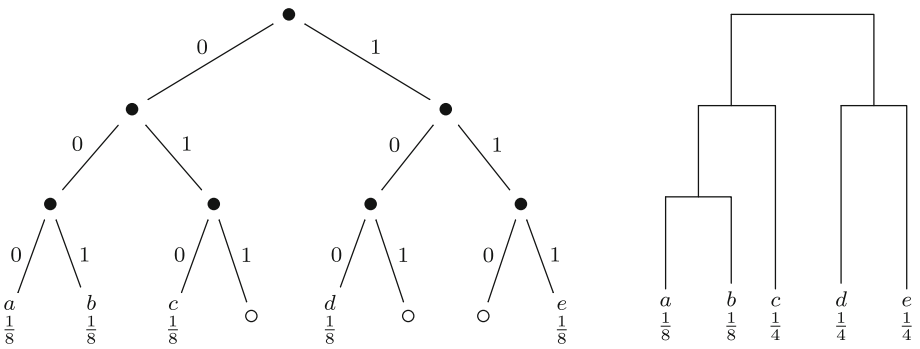

Fig. 2 Left: A 2-adic encoding with Haar measure values assigned to data points. Right: The corresponding dendrogram of the whole dataset with equity probability measure values assigned to data points

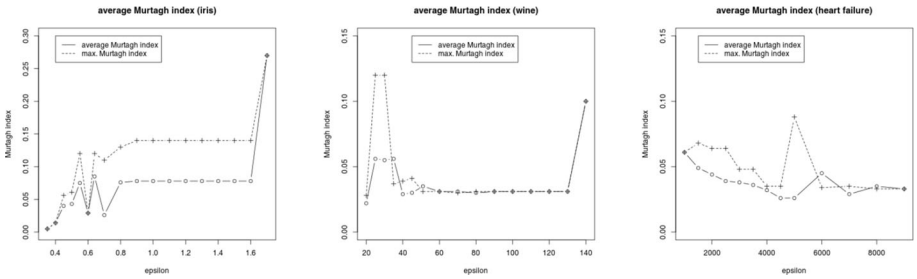

Fig. 3 Average and maximal Murtagh ultrametricity index values across the larger clusters

For $x \in X$ define

$$C_\epsilon^\delta(x) = \text{the connected comp. of } \Gamma_\epsilon^\delta \text{ containing } C(x) \in \mathcal{C}(\Gamma_\epsilon)\,,$$

where $C(x) \in \mathcal{C}(\Gamma_\epsilon)$ is the connected component of Γ_ϵ containing $x \in X$. The local ϵ-δ-genus is the function

$$g_\epsilon^\delta \colon X \to \mathbb{N},\ x \mapsto b_1(C_\epsilon^\delta(x))\,,$$

where C_ϵ^δ is viewed as a subgraph of Γ_ϵ^δ.

Lemma 2.5 *It holds true that*

$$0 \le g_\epsilon^\delta(x) \le \frac{1}{2}\left|C_\epsilon^\delta(x)\right|^2 - \frac{3}{2}\left|C_\epsilon^\delta(x)\right| + 1$$

for $\delta \ge \epsilon > 0$. (X, d) is ultrametric, if and only if for all $\delta \ge \epsilon > 0$, the right-hand side is an equality.

Proof The ϵ-δ-genus is non-negative and is maximal, when $C_\epsilon^\delta(x)$ is a complete graph, in which case the first Betti number equals the right-hand side of the asserted inequality. The last statement now follows immediately from Bradley (2016, Lem. 2.2) applied to the quotient metric on the set $\mathcal{C}(\Gamma_\epsilon)$ induced by d. □

Such a connected component $C = C_\epsilon^\delta(x)$ for $x \in X$ can be viewed as a coarse graph structure on the connected components of Γ_ϵ. Now, replacing, as in the previous subsection, the distance d, restricted to each element of $\mathcal{C}(\Gamma_\epsilon)$ which is contained in C, with an ultrametric, leads to a local tree structure on C. Now, the previous subsection tells us that locally, there

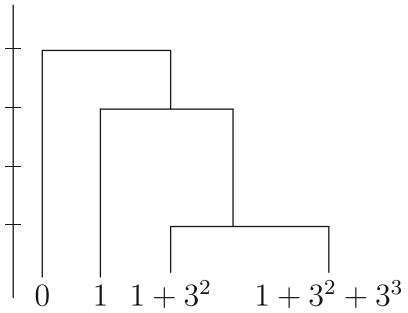

Fig. 4 A 3-adic dendrogram of the set $\left\{0, 1, 1+3^2, 1+3^2+3^3\right\}$

is a Radon measure $\nu(x)$ coming from a p-adic differential 1-form ω_ϵ which is algebraic. The local pieces can now be "glued" to a covering of the p-adic points of a Mumford curve $\mathcal{C}_\epsilon$ minus the zeros of the regular algebraic differential 1-form ω_ϵ, whose genus equals the first Betti number of C. Notice that the gluing process takes place beyond the mere points whose coordinates are in $\mathbb{Q}_p$, in the category of p-adic rigid analytic spaces, as laid out, e.g., in Fresnel and van der Put 2004, Ch. 5). The method from the previous subsection fills a gap of Bradley and Ledezma (2024, §3.3) by explicitly constructing the equity measure on the Mumford curve, which according to Bradley and Ledezma (2024, Lem. 3.9) comes from an algebraic differential 1-form ω_ϵ.

Hence, apart from the local ϵ-δ-genus, there is also the equity measure $|\omega_\epsilon|_p$ on each connected component of Γ_ϵ^δ as a further set of invariant of the dataset (X, d). As another invariant, we suggest also the minimal value δ for which Γ_ϵ^δ is connected, together with the now global ϵ-δ-genus and equity measure for this δ.

Example 2.6 In the toy example of Fig. 1, the coarse graphs Γ_1^3 (middle) and Γ_2^3 (right) are depicted. Observe that the invariants g_ϵ^δ are constant in these cases:

$$g_1^3 = 1, \quad g_2^3 = 1\,,$$

but also

$$g_1^2 = 0$$

is constant, since all three connected components of Γ_1^2 do not contain any cycles.

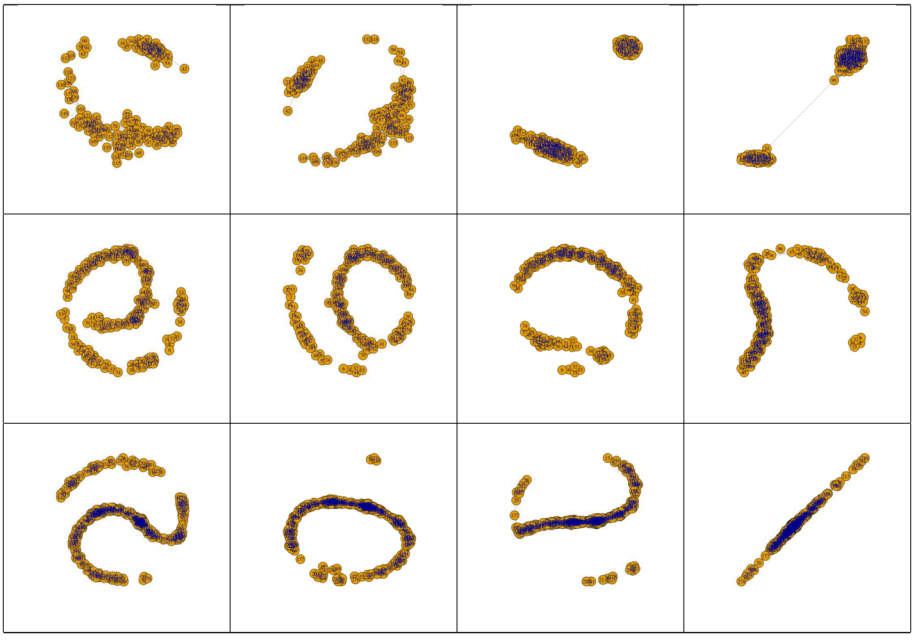

Fig. 5 Top row: Vietoris-Rips graphs Γ_ϵ (iris) for $\epsilon = 0.64, 0.7, 1.640, \epsilon = 1.650$ (from left to right). Middle row: Γ_ϵ (wine) for $\epsilon = 36.8, 40, 51, 70$ (from left to right), having 5, 4, 3, 2 clusters, respectively. Bottom row: Γ_ϵ (heart failure) for $\epsilon = 10,000, 15,000, 20,000, 40,000$, respectively. All rows: Only the non-singleton connected components are shown

3 Experiments

The following datasets were investigated: the iris dataset (Fisher, 1936), wine (Aeberhard & Forina, 1992), and heart failure (Chicco & Jurman, 2020). The Vietoris-Rips graph Γ_ϵ was calculated in R (R Core Team, 2021) for ϵ between 0 and the maximal distance value, using the TDApplied package (Fasy et al., 2015). A visual inspection of the number of connected components (i.e., clusters) of Γ_ϵ^δ via their barcodes leads to identifying ϵ-values with a relatively small number of clusters. Average and maximal Murtagh ultrametricity index values across larger clusters were computed, cf. Fig. 3. Dendrograms are p-adically encoded like in Fig. 4 in order to compute invariants.

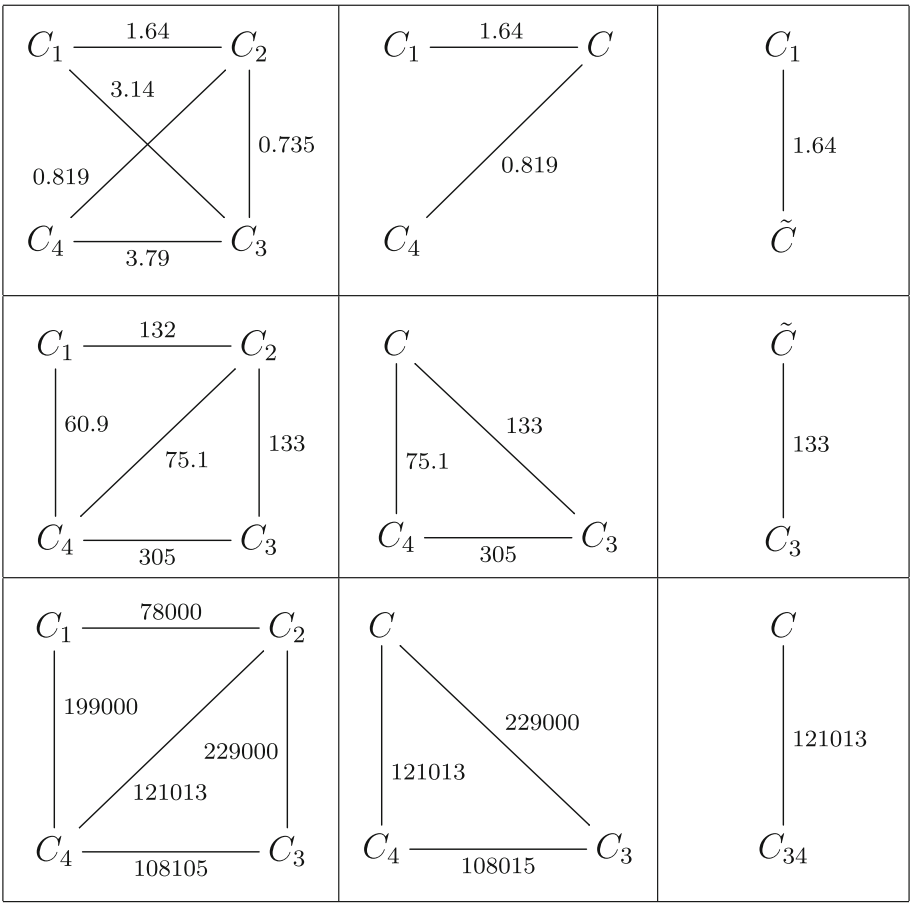

Fig. 6 Evolution of small coarse graphs Γ_ϵ^δ starting at genus 2 w.r.t. ϵ. Top row: The coarse graphs Γ_ϵ^δ (iris) for $(\epsilon, \delta) = (0.7, 4), (0.8, 4), (0.9, 4)$ (from left to right) jump immediately down to genus zero. Middle row: The coarse graphs Γ_ϵ^δ (wine) for $(\epsilon, \delta) = (55, 305), (70, 305), (80, 305)$ (from left to right) have an intermediate graph of genus 1. Bottom row: The coarse graphs Γ_ϵ^δ (heart failure) for $(\epsilon, \delta) = (50,000, 23,000), (95,000, 230,000), (120,000, 230,000)$ (from left to right) have an intermediate graph of genus 1

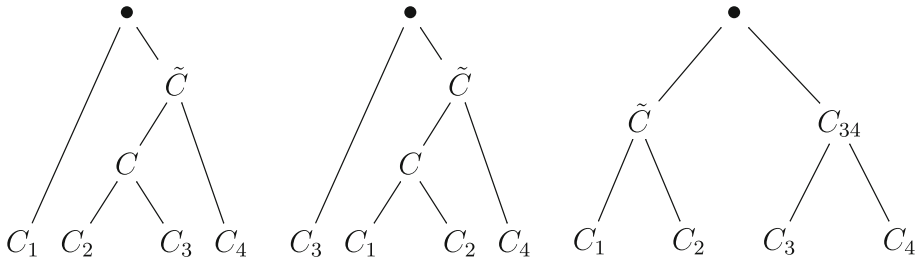

Fig. 7 The cluster hierarchy of the graphs in Fig. 6. From left to right: iris, wine, heart failure

Figure 5 shows Vietoris-Rips graphs for four selected values of ϵ, where singleton clusters are removed from the visualisation. In the iris data case, the graph $\Gamma_{1.650}$ is connected, whereas in the other two cases, the shown graphs are not connected.

Figure 6 shows the two quotient graphs Γ_ϵ^δ for values of (ϵ, δ) such that initially, it is connected with four vertices, and the corresponding Mumford curve has genus 2. Increasing ϵ then either leads to a direct jump down to genus 0 in the iris data case or a descent into genus 0 via an intermediate Mumford curve of genus 1. In the iris dataset case, this means that it contains cycles at a relatively coarse level which is finer than the usual classification into three clusters. This indicates that the three-cluster classification hides some topological structure which only becomes revealed after subdividing the largest cluster into three subclusters. In the other two datasets, this hidden topological structure begins to reveal itself already at a coarser level. The hierarchical structure of the clusters for the corresponding ϵ-values is given in Fig. 7. Tables 1, 2, and 3 show the distances between the clusters of the respective largest coarse graphs of Fig. 6.

Our choice of an ultrametricity is biased according to Murtagh (2004, §3.3) who introduced his ultrametricity index also because of the chaining effect problem of the ultrametricity index from Rammal et al. (1986) based on the subdominant ultrametric, and applied it also to the iris dataset. The Murtagh ultrametricity index values of the not-too-small clusters of the three datasets studied here were calculated using the method below. Figure 3 shows the averages and the maximal values on clusters having at least size 6, whereby ignoring those clusters with Murtagh index precisely zero. The rationale behind this is that a reasonable multi-variate dataset should have at least some ultrametric triangles, and if a dataset is too small, then having relatively few such triangles immediately means that there are none. It was observed that for certain values of ϵ, there is an occasional cluster with exceptionally high ultrametricity value. In the iris and wine datasets, it even happens that the last singleton to be captured into the one big cluster for large ϵ is at a large distance from the other points, and this leads to a very

Table 1 Cluster distances at $\epsilon = 0.7$ (iris)

$\epsilon = 0.7$	C_1	C_2	C_3	C_4
C_1	0	1.64	3.14	5.46
C_2	1.64	0	0.735	0.819
C_3	3.14	0.835	0	3.79
C_4	5.46	0.819	3.79	0

Table 2 Cluster distances at $\epsilon = 55$ (wine)

$\epsilon = 55$	C_1	C_2	C_3	C_4
C_1	0	132	360	60.9
C_2	132	0	133	75.1
C_3	360	133	0	305
C_4	60.9	75.1	305	0

large amount of additional almost ultrametric triangles, so that the Murtagh index value has a sharp jump upwards at the end. In any case, variation in ultrametricity across the clusters occurs for all the datasets. There is no immediate explanation for the fluctuation patterns of the values.

The method of Murtagh (2004, §3.3) of calculating the Murtagh ultrametricity index consists in counting almost ultrametric triangles as follows:

1. Randomly sample the coordinates for triples of three distinct points.
2. Check for possible degenerate triangles and exclude these.
3. The cosine of the angle facing a side of length x is as follows:

$$\frac{y^2 + z^2 - x^2}{2yz},$$

   where y and z are the other side lengths.
4. For the two other angles, seek an angular difference of at most 2° (0.03490656 radians).

Murtagh's ultrametricity index is then the fraction α of such almost ultrametric triangles.

The 3-adic Radon measure leading to the equity measure on one cluster is now computed as an exemplary study of the iris dataset. More precisely, it is a cluster consisting of four points in the graph $\Gamma_{0.64}$, and it is called here $C_{0.64,3}$. The subdominant ultrametric (computed using the method of Rammal et al. 1986) of cluster $C_{0.64,3}$ can be depicted as the dendrogram w.r.t. single-linkage clustering, and is shown in Fig. 8 (top left). As an alternative method, the Ward hierarchical classification was used. The dendrograms for both methods applied to more clusters of the three datasets are shown in Fig. 8 and Fig. 9. In cluster $C_{0.64,3}$, the Ward classification method leads to point 61 being closer to the set {58, 94} than 99, which is not the case for the single-linkage dendrogram.

Since the unlabelled trees corresponding to the different ultrametrics are isomorphic, it is possible to obtain a 3-adic equity measure for all of them as follows: assign to each point of $C_{0.64,3}$ a 3-adic ball of radius 3^{-2}, centred in

$$0, \quad 1, \quad 1 + 3^2, \quad 1 + 3^2 + 3^3,$$

Table 3 Cluster distances at $\epsilon = 50,000$ (heart failures)

$\epsilon = 50000$	C_1	C_2	C_3	C_4
C_1	0	78,000	307,000	199,000
C_2	78,000	0	229,000	121,013
C_3	307,000	229,000	0	108,105
C_4	199,000	121,013	108,105	0

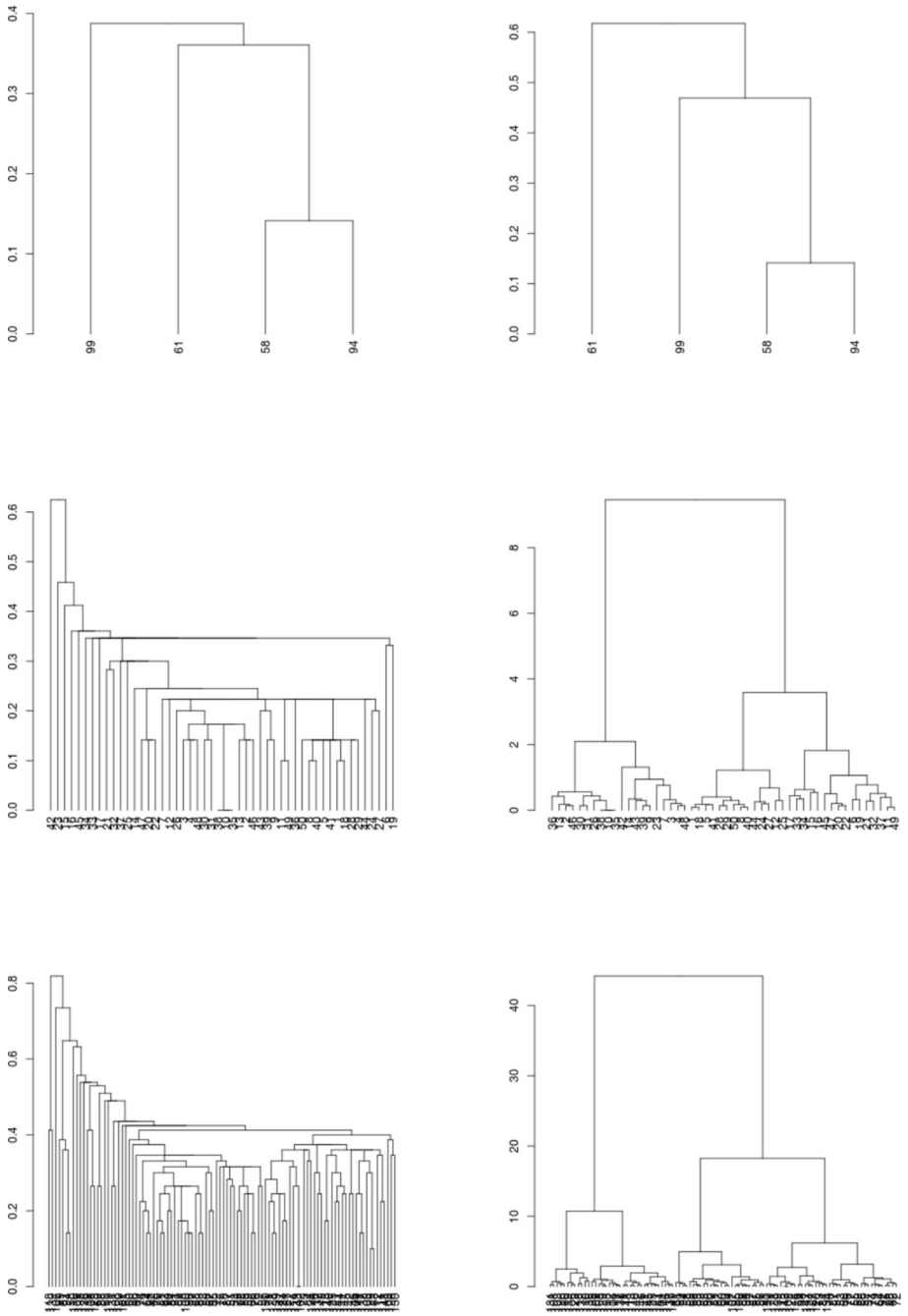

Fig. 8 Dendrograms of the iris data clusters $C_{0.64,3}$ in $\Gamma_{0.64}$, C_1 in $\Gamma_{0.9}$, and $\tilde{C}$ in $\Gamma_{0.9}$ using single-linkage (left column), and Ward (right column) hierarchical classification methods, respectively

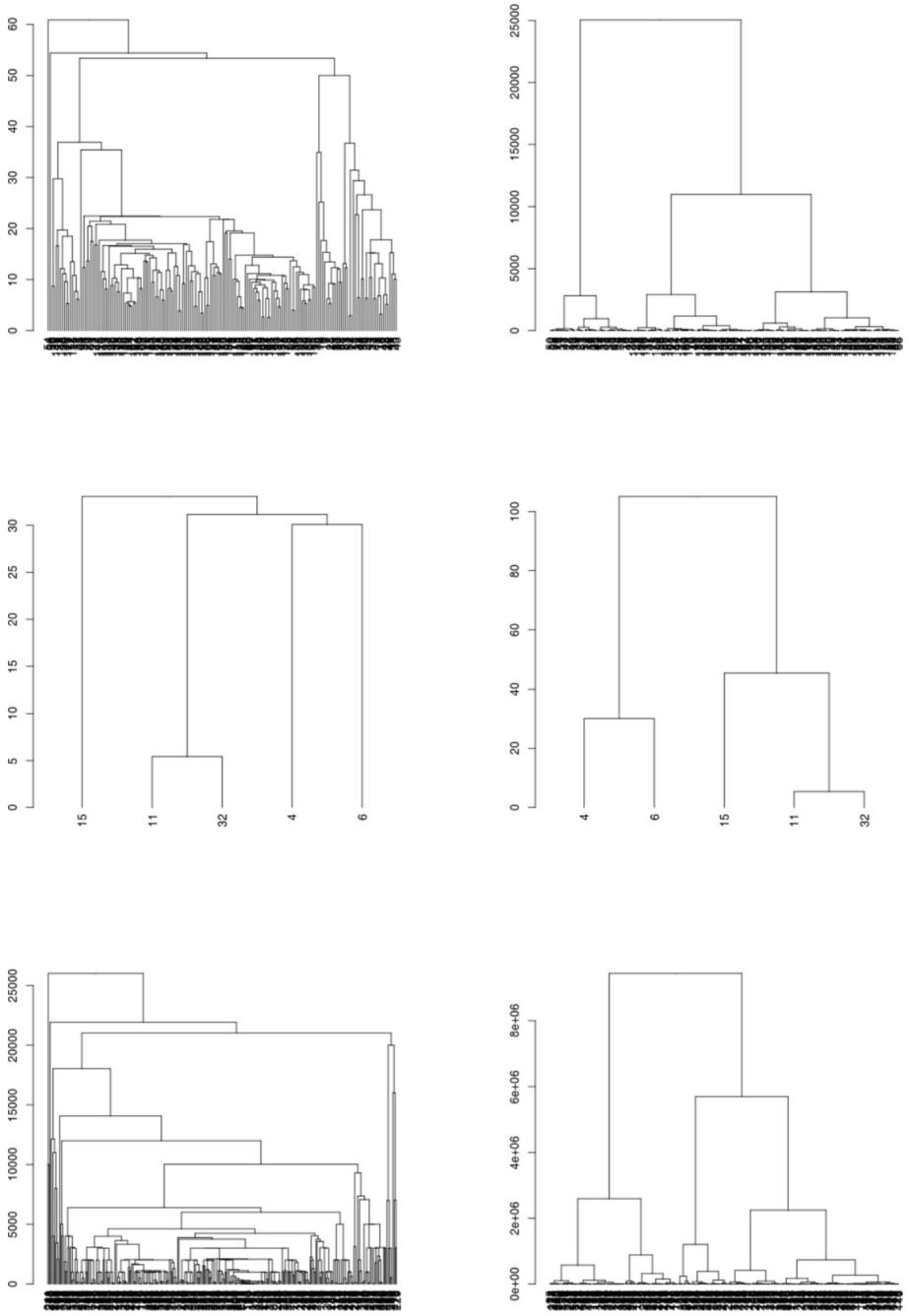

Fig. 9 Dendrograms of the clusters C in Γ_{70} (wine), C_4 in Γ_{80} (wine), and C_1 in Γ_{50000} (heart failures), using single-linkage (left column) and Ward (right column) hierarchical classification methods, respectively

Table 4 3-adic encodings of the elements of cluster $C_{0.64,3}$ in $\Gamma_{0.64}$ of the iris dataset for the used hierarchical classifications

3-adic number	0	1	$1+3^2$	$1+3^2+2^3$
Single-linkage	99	61	58	94
Ward	61	99	58	94

a set which has the 3-adic tree structure of Fig. 8. The equity measure leads to the assignment

$$0 \mapsto \frac{1}{2}, \quad 1 \mapsto \frac{1}{4}, \quad 1+3^2 \mapsto \frac{1}{8}, \quad 1+3^2+3^3 \mapsto \frac{1}{8},$$

having a dendrogram as in Fig. 4.

This leads to the 3-adic interpolation problem

$$f(0)=1, \quad f(1)=3, \quad f(1+3^2)=3^2, \quad f(1+3^2+3^3)=3^2,$$

with the solution

$$f(X)=\frac{31}{9990}X^3-\frac{1683}{9990}X^2+\frac{10811}{4995}X+1 \in \mathbb{Q}_3[X]$$

for the new measure

$$|\omega(x)|_3 = |f(x)|_3 \ |dx|_3$$

approximating the equity measure for $x \in C_{0.64,3}$. Now, make the assignment as in Table 4.

4 Conclusion

Hierarchical classification aims at finding a suitable ultrametric in a dataset which resembles as closely as possible the metric or similarity structure inherent to the given data. The idea behind this present work is that an actual inherent similarity structure in data could possibly only be local. Hence, following this idea, it should be better to find a suitable locally ultrametric approximation to given metric data instead of a global ultrametric. This then opens a way for revealing hidden inter-cluster topology in data, which is impossible to detect with a global ultrametric, i.e., by hierarchically classifying the whole dataset. The findings suggest that this idea is justified. In particular, the variation in local ultrametricity indicates that a dataset in general cannot be expected to be homogeneous with respect to its inherent hierarchical structure.

In the theoretical part of this work, a local ultrametric is associated with a finite metric space (given the meaning of a dataset) via its Vietoris-Rips graph. The equity measure was defined on the local ultrametric parts of the space which, by the result of recent previous work, can be seen as coming from a differential 1-form on a p-adic Mumford curve. These can be viewed as compact algebraic p-adic manifolds, and this approach leads to new invariants for finite metric data by taking a double filtration with ϵ- and δ-balls in the finite metric. Like in the case of topological or multi-topological data, this means a p-adic manifold interpretation of data via the indexing structure defined for accessing purposes through the subdominant ultrametric of the graph distance proposed in Bradley and Ledezma (2024), generalising the approach of Bradley and Jahn (2022) to higher dimensional structures, and that actually any ultrametric can be used.

In the experimental part of this work, it is conceivable from the findings with the iris dataset that local ultrametricity could be increased within some clusters of the Vietoris-Rips

graphs in comparison with the ultrametricity of the whole dataset, whereas in other clusters, it could be significantly lower. This resembles Simpson's paradox in statistics. Even if this may not be always the case for general datasets, this nevertheless provides a means for classifying different types of datasets. Exemplary, Vietoris-Rips graphs were taken, and some values of the new invariants were calculated. The findings suggest that the double filtration approach can reveal more inherent topological properties of data in their ultrametric approximation via p-adic encoding. This suggests that in the future, a hierarchical or p-adic version of manifold learning could emerge from further investigations in this direction, which is appealing because of its potential for a reduced computational complexity.

Potential applications are those situations in which the hierarchical classification of whole datasets does not make sense, either for reasons inherent to the dataset or computationally due to the data size. Ongoing research in this direction is motivated by applications in geoinformatics where temperature flows are to be simulated on large city models on distributed processing systems.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability This research has used publicly available data only. The R code is available under https://github.com/pebradley/localUltrametricity and is published under a GNU General Public License.

References

Aeberhard, S., & Forina, M. (1992). Wine [dataset]. UCI Machine Learning Repository.

Benois-Pineau, J., & Khrennikov, A. (2010). Significance delta reasoning with p-adic neural networks: Application to shot change detection in video. *The Computer Journal, 53*(4), 417–431.

Bradley, P. (2006). Degenerating families of dendrograms. *Journal of Classification, 25*(1), 27–42.

Bradley, P. (2010). Mumford dendrograms. *The Computer Journal, 53*(4), 393–404.

Bradley, P. (2016). Ultrametricity indices for the Euclidean and Boolean hypercubes. *p-Adic Numbers, Ultrametric Analysis, and Applications, 8*(4), 298–311.

Bradley, P. (2017). Finding ultrametricity in data using topology. *Journal of Classification, 34*, 76–84.

Bradley, P. (2019). On the logistic behaviour of the topological ultrametricity of data. *Journal of Classification, 36*, 266–272.

Bradley, P., & Jahn, M. (2022). On the behaviour of p-adic scaled space filling curve indices for high-dimensional data. *The Computer Journal, 65*(2), 310–330.

Bradley, P., & Ledezma, A. (2024). Approximating diffusion on finite multi-topology systems using ultrametrics. arXiv:2411.00806 [cs.DM].

Chicco, D., & Jurman, G. (2020). Heart failure clinical records [dataset]. UCI Machine Learning Repository.

Dragovich, B., Khrennikov, A., Kozyrev, S., & Mišić, N. (2021). p-adic mathematics and theoretical biology. *BioSystems, 199*, 104288.

Fasy, B., Kim, J., Lecci, F., & Maria, C. (2015). Introduction to the R package tda. hal-01113028.

Fisher, R. (1936). Iris [dataset]. UCI Machine Learning Repository.

Fresnel, J., & van der Put, M. (2004). *Rigid analytic geometry and its applications*. Progress in Mathematics. Birkhäuser, Boston.

Gerritzen, L., & van der Put, M. (1980). *Schottky groups and Mumford curves* Lecture Notes in Mathematics (Vol. 817). Heidelberg, New York: Springer.

Murtagh, F. (2004). On ultrametricity, data coding, and computation. *Journal of Classification, 21*, 167–184.

Murtagh, F. (2016). Sparse p-adic data coding for computationally efficient and effective big data analytics. *p-Adic Numbers Ultrametric Analysis and Applications, 8*, 236–247.

R Core Team. (2021). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rammal, R., Toulouse, G., & Virasoro, M. (1986). Ultrametricity for physicists. *Reviews of Modern Physics, 58*(3), 765–788.

Shor, O., Glik, A., Yaniv-Rosenfeld, A., Valevski, A., Weizman, A., Khrennikov, A., & Benninger, F. (2021). EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline. *PLOS ONE, 16*(8), e0255529.

Vietoris, L. (1927). Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen, 97*(1), 454–472.

Vladimirov, V., Volovich, I., & Zelenov, E. (1994). *p-adic analysis and mathematical physics*. Series on Soviet and East European Mathematics, 1. World Scientific Publishing Co., Inc., River Edge, NJ.

Zomorodian, A. (2010). Fast construction of the Vietoris-Rips complex. *Computers & Graphics, 34*(3), 263–271.

Zubarev, A. (2014). On stochastic generation of ultrametrics in high-dimensional Euclidean spaces. *p-Adic Numbers, Ultrametric Analysis, and Applications, 6*(2), 155–165.