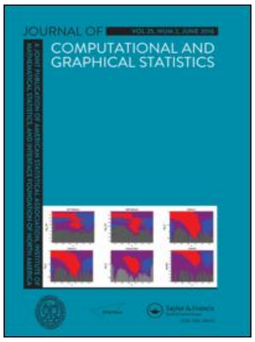
Thomas Rusch and Kurt Hornik and Patrick Mair

Assessing and quantifying clusteredness: The OPTICS Cordillera



Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Assessing and quantifying clusteredness: The OPTICS Cordillera

Thomas Rusch, Kurt Hornik & Patrick Mair

To cite this article: Thomas Rusch, Kurt Hornik & Patrick Mair (2017): Assessing and quantifying clusteredness: The OPTICS Cordillera, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2017.1349664

To link to this article: http://dx.doi.org/10.1080/10618600.2017.1349664

Assessing and quantifying clusteredness: The OPTICS Cordillera

Thomas Rusch*
Competence Center for Empirical Research Methods
WU (Vienna University of Economics and Business)
Austria
and
Kurt Hornik
Institute for Statistics and Mathematics
WU (Vienna University of Economics and Business)
Austria
and
Patrick Mair
Department of Psychology
Harvard University
Cambridge, MA

June 22, 2017

*The authors gratefully acknowledge the associate editor and three anonymous reviewers who helped much in improving the paper.

Abstract

This paper provides a framework for assessing and quantifying "clusteredness" of a data representation. Clusteredness is a global univariate property defined as a layout diverging from equidistance of points to the closest neighbouring point set. The OPTICS algorithm encodes the global clusteredness as a pair of clusteredness-representative distances and an algorithmic ordering. We use this to construct an index for quantification of clusteredness, coined the OPTICS Cordillera, as the norm of subsequent differences over the pair. We provide lower and upper bounds and a normalization for the index. We show the index captures important aspects of clusteredness such as cluster compactness, cluster separation and number of clusters simultaneously. The index can be used as a goodness-of-clusteredness statistic, as a function over a grid or to compare different representations. For illustration, we apply our suggestion to dimensionality reduced 2D representations of Californian counties with respect to 48 climate change related variables. Online supplementary material is available (including an R package, the data and additional mathematical details).

Keywords: index, cluster analysis, dimensionality reduction, perception

1 Introduction and Motivation

Representation of a data matrix in $\mathbb{R}^m$ is an integral part of exploratory data analysis. Often the representation is interpreted by quantification of structure (e.g., by a correlation) and/or by inspection of the visual appearance. One type of structure frequently of interest is whether and how the data points are arranged in discrete groups (clusters). We call this "clusteredness".

Clusteredness is a somewhat elusive concept. It has been discussed to some extent in the literature (e.g., in Greenacre, 2011), but its definition remains vague. Clusteredness is often assessed visually from how clustered the points in a representation appear. This process is unclear and intransparent, depending largely on the observer and precluding a sensible, replicable quantification. It is also limited to (series of) representations in $\mathbb{R}^2$ or $\mathbb{R}^3$.

For illustration and to motivate the paper, the last row of Figure 1 shows six different 2D scatterplots depicting the same data: a random subset of 100 cases of the handwritten digits 1-4 from Alimoglu (1996) (dimensionality reduced). The representations clearly differ in how clustered they appear.

We asked a diverse set of 24 subjects (see supplement) to rank order the plots according to the perceived clusteredness of the results, solely instructing that all plots show exactly the same data (including the same number of data points).The subject's ranking patterns are given as a parallel coordinate plot (jittered) in the top row of Figure 1.

The picture is striking: The 24 subjects made 20 different rankings. Clusteredness of the plots is judged very differently—at most three subjects agreed on a common ranking. There is little overall consensus and aggregation into a common ranking is not straightforward. Quantification of the degree of clusteredness of the representations was reported as very difficult. It seems likely that the different rankings stem from the observers having different implicit views of clusteredness and how to judge it. So, while the approach of visually interpreting clusteredness data representations is common, it appears to be highly subjective. The aim of this paper is to provide a clearly defined way to assess, quantify and interpret the clusteredness of a data matrix in $\mathbb{R}^m$.

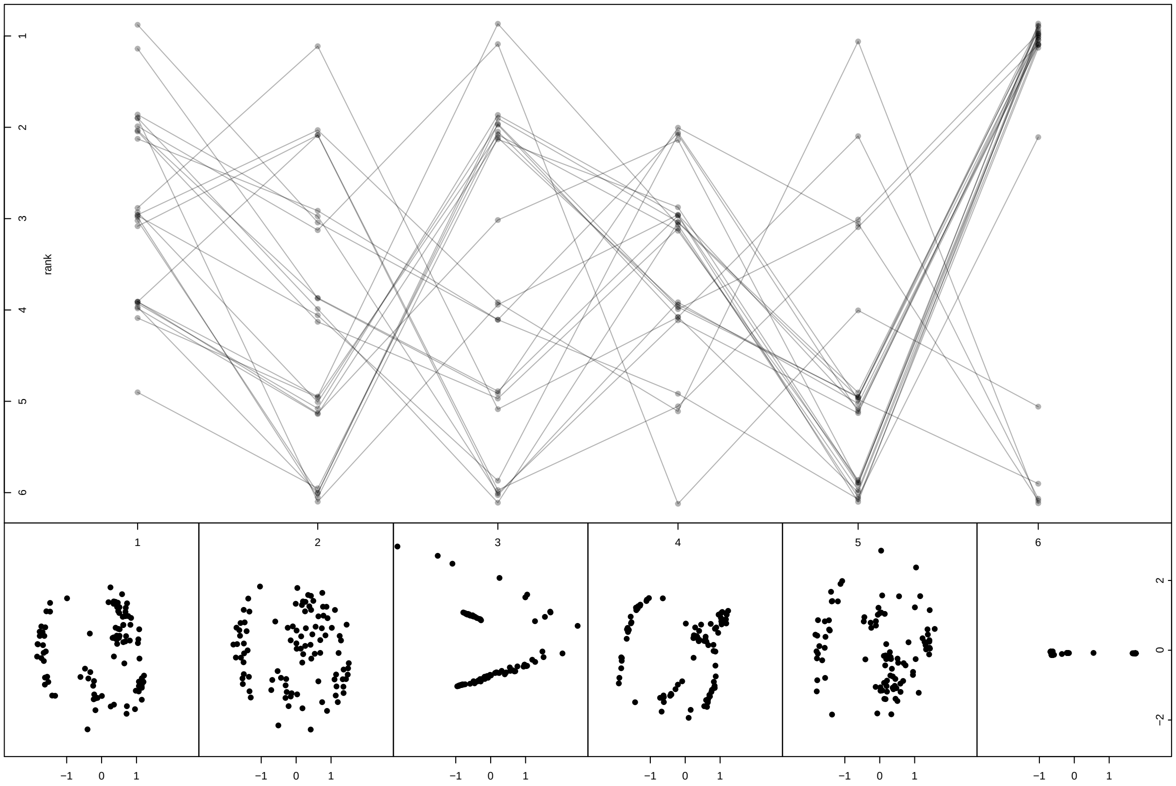


Figure 1: Parallel coordinate plot of ranking patterns of the 2D plots in the bottom row with respect to the perceived clusteredness for $n = 24$ subjects. Jittering was applied to points to improve readability.

The paper is organized as follows: First, building upon the density-based clustering concept underlying Ester et al. (1996) and its extensions, in Section 2 we define and formalize a density- and distance-based notion of clusteredness of a data representation as a continuum of appearances of a representation with no clusteredness and maximal clusteredness as endpoints. Second, we discuss aspects of clusteredness relevant to determine the position on the continuum. These aspects relate to (i) whether a specific number of objects accumulate close by each other, (ii) how densely the objects accumulate, (iii) how separated the accumulations are, (iv) the number of accumulations and (v) the spread of objects in $\mathbb{R}^m$. Third, in Section 3 we propose a univariate measure for quantifying global clusteredness, the Cordillera, which assesses how much clusteredness one finds in a data representation. It is based on a clusteredness-representative algorithmic ordering and clusteredness-representative distances ("reachabilities"). We suggest a specific instance of the Cordillera utilizing the OPTICS (Ordering Points To Identify The Clustering Structure; Ankerst et al., 1999) algorithm for obtaining the algorithmic ordering and reachabilities, which fits neatly into the distance-density based framework and has the properties of making only weak assumptions about the object arrangement in the representation. We call this instance the OPTICS Cordillera, and we give results on its behavior. In Section 4 we illustrate the practical usage of our proposal on dimension reduction results of a data set on climate change related natural hazards for Californian counties. We finish with some final remarks in Section 6.

2 Clusteredness

In this section we describe and formalize a conceptual framework of clusteredness which captures the accumulation tendency of objects based on the minimum number of objects comprising an accumulation (k) and the density of objects within a radius up to $\epsilon_{\max}$. In Section 3 we use this framework to develop an index that quantifies a representation's clusteredness.

For notation, let $x_1, \dots, x_N \in \mathbb{R}^m$. The $x_1, \dots, x_N$ (points or objects) are row vectors of the data representation as data matrix X. Let $d_{ij} = d(x_i, x_j)$ denote a distance between the observations x_i and x_j, most naturally induced by a norm; typically the p-norm distance

$d_{ij} = ||x_i - x_j||_p$ with $p \geq 1$.

2.1 Distance-density based Accumulations

We adopt the distance-density based concept of clusters suggested in Ester et al. (1996). This concept builds in parts on ideas summarized in Jain and Dubes (1988). The underlying notion is that a cluster is an arbitrarily shaped accumulation of points with a given density. Density is here related to a counting measure on the set of points for a given ϵ-neighbourhood. Note that in this definition the shape of the neighbourhood depends on the distance function chosen.

The density of points in an accumulation is assumed to be higher than the density of points between accumulations. In essence, areas of high density define accumulations which are separated by low density areas. Areas of noise may also exist; they are characterized by a density that is lower then the density in any accumulation. An important part in this concept is that the density in an area has to exceed a certain threshold, which means that the accumulation must exceed a given number of points k. This k is the minimum number of points that must comprise an accumulation (i.e., $2 \leq k < N$) and determines the density threshold.

Let $N_\epsilon(x_i) = \{x_j : d_{ij} < \epsilon\}$ be the set of neighbouring objects to x_i within a radius of ϵ, including x_i itself (ϵ-neighbourhood of x_i). Subsequently we always include the point x_i when counting, so the 1-st nearest neighbour to point x_i is point x_i itself. Let $S_{k,\epsilon}(x_i)$ be the subset of $N_\epsilon(x_i)$ that contains the $k-$th nearest neighbouring point(s) to and including x_i, $S_{k,\epsilon}(x_i) \subseteq N_\epsilon(x_i)$. If $\text{card}\,(N_\epsilon(x_i)) < k$, then $S_{k,\epsilon}(x_i) = \emptyset$. Let a point x_j be called *directly density reachable* from point x_i if $x_j \in N_\epsilon(x_i)$ and $\text{card}\,(N_\epsilon(x_i)) \geq k$; x_i is then called a *core point*. Further, let x_j be called *density reachable* from point x_i if there is a chain of directly density reachable points from x_i to x_j. Let a point x_j be called *density connected* to x_i if there is a point x_k from which both x_i and x_j are density reachable. Then we define an *accumulation* of points or objects as (cf. Ester et al., 1996):

Definition 1 (Accumulation)**.** Given ϵ and k an accumulation is a (non-empty) set C of at least k objects satisfying

$$\forall\ x_i, x_j : \text{If } x_i \in C \text{ and } x_j \text{ is density reachable from } x_i \implies x_j \in C$$
$$\forall\ x_i, x_j \in C : x_i \text{ and } x_j \text{ are density connected}$$

Every object not in any accumulation is “noise”. A specific accumulation of k objects of which x_i is a member we denote by $C_{k,\epsilon}(x_i)$.

An accumulation is therefore a set of density-connected points which is maximal with respect to density reachability for given k, ϵ (the first part of Definition 1). An accumulation is judged solely by the distances between the objects and the density of the objects in the accumulation. The particular shape of the accumulations remains unspecified.

Definition 1 also means that for given k and ϵ any object is either an element of only one accumulation or “noise”, which makes it a definition suitable for partitional clustering.

2.2 Distance-density based Clusteredness

We understand clusteredness as a unidimensional representation of the global clustering structure. It is the global tendency of X to have points arranged in an unspecified number of appreciable accumulations of size $\geq k$. We are interested if and how accumulations are present in X rather than the specific accumulations themselves.

An important part of clusteredness as a global property is that for given k but varying ϵ, any object can be an element of many, possibly nested accumulations or is “noise”. Particularly, given k, every accumulation with respect to ϵ_l is a subset of an accumulation with respect to ϵ_m if $\epsilon_m > \epsilon_l$. To capture the global clustering structure (overall clusteredness) one needs to simultaneously characterize the presence of any number of accumulations for all ϵ up to an $\epsilon_{\max} = \sup \epsilon$. In some sense this is the hierarchical clustering extension of the distance-density based partitional clustering idea from above.

The simultaneous characterization of the global clustering structure is typically not encoded univariately but, say, in a dendrogram. A specific algorithm that gives such an encoding in the framework reviewed above will be discussed in Section 3.1. We are concerned with obtaining a sensible, unidimensional characterization of clusteredness from an encoding of the global distance-density based clustering structure. We first define the unidimensional continuum of clusteredness.

The Clusteredness Continuum. We define clusteredness as the unidimensional continuum representing the global distance-density based arrangement of the objects in X in accumulations of size $\geq k$ for varying ϵ. The two possible extremes are no clusteredness ("unclusteredness") and maximal clusteredness. The degree of clusteredness of X is the tendency of points in X to be arranged in accumulations, i.e., the divergence of X from an unclusteredness layout, which is maximal for maximal clusteredness.

In this framework, unclusteredness is characterized by the situation that all points have $\geq k-1$ neighbours at a constant distance c. $k = 2$ gives the most extreme unclustered layouts and a typical example is when all density-reachable points for $\epsilon = c$ fall onto a regular tessellation or a lattice in $\mathbb{R}^m$.

The extreme of a maximally clustered arrangement is when for the maximum possible number of "groups" of points (i.e., the maximum possible number of accumulations plus one optional group of points that are not in any accumulation), the k observations comprising an accumulation coincide exactly (no distance to each other) and all the positions of accumulations are at a constant distance $d_{\max}$ away from the closest neighbouring accumulation(s) ($d_{\max}$ also being the minimum distance between any two points from two closest neighbouring accumulations), so there is equidistance among the closest neighbouring accumulations.

We can state this more formally:

Definition 2 (No clusteredness)**.** For given $2 \leq k < N$ and (sufficiently large) $\epsilon_{\max} > 0$ so that $S_{k,\epsilon_{\max}}(x_i) \neq \emptyset \; \forall \; x_i$, no clusteredness ("unclusteredness") is given if there $\exists \; c, 0 < c < \infty$, so that $\forall \; x_i : d(x_i, x_j) = c$ if $x_j \in S_{k,\epsilon_{\max}}(x_i), x_i \neq x_j$.

Definition 3 (Maximal clusteredness)**.** For given $2 \leq k < N$ and (sufficiently large) $\epsilon_{\max} > 0$, let n denote the maximum number of possible groups of objects that can be formed for given k and N, comprising the number of the $\lfloor N/k \rfloor$ accumulations and possibly an additional group of $N - \lfloor N/k \rfloor k$ points that cannot form an accumulation, so $n = \lceil N/k \rceil$ groups. Maximal clusteredness is an arrangement with $n = \lceil N/k \rceil$ groups where for all k

objects in an accumulation it means that $x_j \in C_{k,\epsilon_{\max}}(x_i) \Leftrightarrow d_{ij} = 0$ and there $\exists\ d_{\max} > 0$ so that $\forall\ i\ \tilde{\delta}_i := \min\{d_{it} : d_{it} > 0\} = d_{\max}$.

The maximal number of groups coincides with the maximal number of accumulations at N/k if $N \equiv 0 \pmod{k}$. Otherwise we have an additional group of $N - \lfloor N/k \rfloor k$ objects that cannot form an accumulation due to the restriction that there must be at least k observations in an accumulation.

The observed degree of clusteredness for X lies on the continuum spanned by the two extremes. To illustrate we use toy examples of 8 labeled data points (see Figure 2). Unclustered arrangements are illustrated in the first row. Maximal clusteredness is illustrated in the bottom right plot where for $N = 8$ objects and $k = 2$ the maximum possible number of groups is $n = 4$. In each of the four accumulations there are $k = 2$ objects coinciding exactly and all four accumulations are equally far away from the closest accumulations. In Figure 2 the plots in between the top row and the bottom right panel show different positions on the continuum in increasingly clustered arrangements with $N = 8$ and $k = 2$; note that the plot in the central panel shows two accumulations of four objects each which with respect to $k = 2$ is less clustered than the subsequent plots in reading order. For $k = 4$ however this plot would be the most clustered of all.

The position on the continuum is related to aspects of the global clustering structure that are all represented in clusteredness. Specifically, clusteredness increases if (i) distances between accumulations increase ("emphasis aspect" or separation), (ii) objects accumulate more densely ("density aspect" or cohesion or compactness), (iii) the number of accumulations increase up to the maximal number ("tally aspect").

We initially derived these aspects in discussion among the authors about desirable properties that should be met for quantifying clusteredness in our framework. Subsequently we found empirical support for the importance of these aspects also in cognitive interviews conducted with subjects who ranked the plots from the introduction. The aspects turned up as themes in the interviews and the plots were ranked according to those themes. This is presented in more detail in the supplementary material.

Our concept of clusteredness is related but distinct from the concept of internal cluster validity as measured for partitional clustering by, e.g., the Silhouette measure (Rousseeuw,

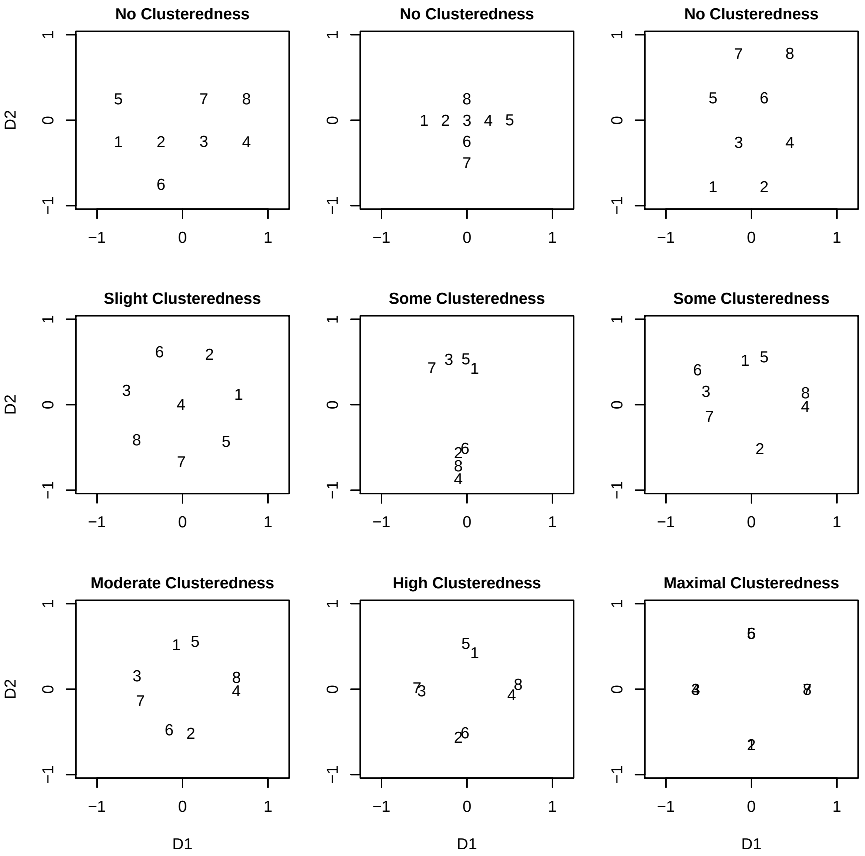


Figure 2: Differently clustered 2D representations of $N = 8$ points. The top row shows cases that are unclustered (Definition 2, $OC' = 0$). The definition of OC' is given in Section 3. The second row shows little to moderately clustered arrangements as measured by the OC' with respect to clusters of at least $k = 2$ points, the bottom panel shows arrangement with higher OC' with respect to $k = 2$. The bottom right plot is maximally clustered for $k = 2$ (Definition 3, $OC' = 1$). All plots show increasing OC' with respect to $k = 2$ in reading order.

1987), prediction strength (Tibshirani and Walther, 2005), the Theoretical Clustering Index (TCI; Huang et al., 2015) and similar indices (see e.g., Liu et al., 2010, for an overview). The main difference lies in the role a specific partitioning plays: In cluster validity, the task is to assess how well a specific clustering represents the objects. For this, a clustering must be found and each observation must be assigned to a cluster. In contrast, clusteredness is a property of the global clustering structure of X and does not entail exclusive cluster assignment of objects.

3 A Clusteredness Index: The OPTICS Cordillera

In this section we propose an index that allows to assess clusteredness as discussed in Section 2. We call our proposal the "OPTICS Cordillera". The index is bounded and owing to the distance-density based framework applicable to a wide range of clustered appearances.

In essence our proposal maps the global clustering structure of X encoded in a hierarchical clustering result to the unidimensional clusteredness continuum, numerically reflecting the degree of clusteredness of X.

The encoding needs to comprise two things: First, a clusteredness-representative algorithmic ordering $R(X) = \{x_{(s)}\}_{s=1,\dots,N}$ which is an ordered set of the original points $x_i, (i = 1, \dots, N)$. So $x_{(1)}$ is the x_i at the first position in $R(X)$. $R(X)$ is obtained by a mapping between the original points and the ordering, $\rho : \{1, \dots, N\} \to \{1, \dots, N\}$, so that $s = \rho(i)$ and $i = \rho^{-1}(s)$. This allows us to refer to the position of point x_i in the ordering $R(X)$ as $x_{\rho(i)}$, or to the point x_i in X for which $x_{(s)} = x_{\rho(i)}$ as $x_{\rho^{-1}(s)}$[1]. Second, an equally-sized associated set of clusteredness-representative distances ("reachability distances" or "reachabilities") $r^*_{(s)} = r^*_{\rho(i)} = r^*_i$ for each point $x_{(s)} = x_{\rho(i)}$ of x_i. "Clusteredness-representative" means that each accumulation in X is sequentially represented in $R(X)$ and that locally maximal $r^*_{(s)}$ in the sequence $R(X)$ characterize separation between accumulations whereas locally minimal $r^*_{(s)}$ in the sequence $R(X)$ characterize compactness in an accumulation. Together the pair $(R(X), \{r^*_{(s)}\}_{s=1,\dots,N})$ encodes the clusteredness structure

[1] The brackets around a subscript only singles out that it is located in the ordering, which we do not need for $\rho(i)$ as that is already the map to the ordering.

in the representation in such a way that the order in $R(X)$ and subsequent reachabilities are representative of clusteredness as defined in the previous section, i.e, that sequentially over the ordering $R(X)$ it holds that if $r^*_{(s)}$ is small then $x_{\rho^{-1}(s)}$ and $x_{\rho^{-1}(s-1)}$ are close together. If $r^*_{(s)}$ is large then $x_{\rho^{-1}(s)}$ is "far away" from $x_{\rho^{-1}(s-1)}$ and also from the other predecessors $x_{\rho^{-1}(s-t)}, t > 1$.

The index, the Cordillera$(R(X), \{r^*_{(s)}\}; q)$ is the q-norm of the finite differences of the $r^*_{(s)}$ over $R(X)$. The name Cordillera comes from an analogy of the plot of reachabilities over the ordering to a mountain range and that the index in a sense measures its raggedness. Below we discuss how to obtain a concrete $R(X)$ and $\{r^*_{(s)}\}$ that is compatible with the distance-density based clusteredness framework and then define the Cordillera for that instance. This instance we coin the OPTICS Cordillera (OC). Note that the Cordillera could be analogously defined in a compatible clusteredness framework for other algorithms yielding an $(R(X), \{r_{(s)}\})$ pair, like CLUES (Wang et al., 2007) or OETICS (Forina et al., 2004).

3.1 Algorithmic Ordering and Reachabilities by OPTICS

The Cordillera depends upon the clusteredness-representative algorithmic ordering or seriation $R(X)$ and the set of clusteredness-representative distances $r^*_{(s)}$. For obtaining a concrete clusteredness-representative ordering-reachability pair we use the OPTICS algorithm (Ordering Points To Identify The Clustering Structure; Ankerst et al., 1999), a sorting algorithm that outputs a linear ordering of objects where each object is associated with a special distance. OPTICS has been developed within the distance-density based framework we used for our clusteredness concept and thus naturally lends itself to substantiate the Cordillera. A central goal that motivated our choice of framework and algorithm was to be as inclusive as possible about the nature of accumulations: allowing for nested accumulations, no need for specific accumulation assignment, no assumption about the exact number of accumulations/observations per accumulation nor about the accumulation shape (beyond the distance) and no reliance on a notion of centroid. The only assumptions we make are that an accumulation must comprise at least k objects and that the objects are density-connected (see Definition 1). OPTICS allows to do that.

Subsequently, we go into detail about the OPTICS Cordillera and its properties. The $OC = 0$ in case of unclusteredness of X as defined in Definition 2 and quantifies how close X is to Definition 3 for given $N, k, \epsilon_{\max}$ and a maximum representative reachability $d_{\max}$. For $k = 2$ and $OC = 0$ implies unclusteredness. The OC is parsimonious with respect to parametrization as only k needs to be specified at all times; it however has additional optional parameters to control runtime and the definition of noise ($\epsilon_{\max}$), "outlier" influence ($\epsilon_{\max}, d_{\max}$) and aggregation of clusteredness aspects (q). We first briefly describe OPTICS and then substantiate the OPTICS Cordillera. We then give lower and upper bounds for it, and discuss properties of the index including its ability to capture the clusteredness aspects.

The OPTICS Algorithm. We only paraphrase the OPTICS algorithm here. In OPTICS two special pairwise distances between points for given k and $\epsilon_{\max}$ are defined (the definitions can be found in Ankerst et al. (1999) or Appendix B): First, the "core distance" c_i which is the distance of a point x_i to its k-th neighbour if $S_{k,\epsilon_{\max}}(x_i) \neq \emptyset$ and undefined otherwise. Second, the "reachability distance" r_{ij} between two points x_i and x_j which is $\max(d_{ij}, c_i)$ if $S_{k,\epsilon_{\max}}(x_i) \neq \emptyset$ and again undefined if $S_{k,\epsilon_{\max}}(x_i) = \emptyset$. The parameter k is mandatory but $\epsilon_{\max}$ is optional and can simply be set "very large" (e.g., $\max d_{ij}$). Smaller $\epsilon_{\max}$ improves runtime of OPTICS and tends to assign more objects as "noise".

Based on these two distances the OPTICS algorithm orders the points to obtain $R(X)$, see Ankerst et al. (1999) or Algorithm 1 in Appendix B. It provides the mapping ρ, which is not expressible in closed form. OPTICS's principle is the following: A point gets visited and the neighbours are recorded. Then its core distance is calculated (if defined, else the next point is used). Then the directly density reachable neighbours get inserted into a priority queue sorted by reachability distance to the closest core point. This queue is iteratively updated for the reachability distance based on the $\epsilon_{\max}$-neighbourhood of the point and the neighbours in the queue. The queue gets processed so that the point with smallest reachability distance is selected, its neighbours get recorded and the core distance gets determined and the reachabilities are updated. If the current point is again a core point, the above is repeated until no unprocessed points are left. Then the closest unprocessed point is selected. During this process each point gets assigned the last updated reachability ("suitable reachability distance") $\tilde{r}_i = \tilde{r}_{\rho(i)}$. The OPTICS algorithm outputs the ordering $R(X)$ together with $\tilde{r}_i = \tilde{r}_{\rho(i)}$.

It is so that if the suitable reachability $\tilde{r}_{\rho(i)}$ for x_i is small then $x_{\rho^{-1}(\rho(i))}$ and $x_{\rho^{-1}(\rho(i)-1)}$ are close together. If it is large then $x_{\rho^{-1}(\rho(i))}$ is far away from $x_{\rho^{-1}(\rho(i)-1)}$ and also from $x_{\rho^{-1}(\rho(i)-t)}, t > 1$. Therefore, points that are subsequent in the ordering $R(X)$ and have small $\tilde{r}_{\rho(i)}$ correspond to accumulations in X, whereas points that are far removed from each other in $R(X)$ or have some large suitable reachability between them appear distant in X.

In the Cordillera we use a winsorized version of $\tilde{r}_i$, r^*_i, that is always defined:

Definition 4 (Representative Reachability)**.** Let $R(X; k, \epsilon_{\max})$ be the OPTICS ordering of X given k and $\epsilon_{\max}$. The representative reachability distance r^*_i for point x_i is

$$r^*_i = r^*_{\rho(i)} = \begin{cases} \min(\tilde{r}_i, d_{\max}) & \text{if } \tilde{r}_i \neq \text{undefined} \\ \min(d_{\max}, \max\{\tilde{r}_i | \tilde{r}_i \neq \text{undefined}\}) & \text{otherwise} \end{cases} \tag{1}$$

Here $d_{\max}$ is the winsorization limit of the maximum possible r^*_i. No winsorization (i.e., $d_{\max} \geq \max \tilde{r}_i$ over the defined $\tilde{r}_i$) may make the index susceptible to outliers. Setting $d_{\max}$ to a threshold below that will winsorize all r^*_i larger than $d_{\max}$ to the value of $d_{\max}$. Note that d_{max} may winsorize the reachabilities for points in the same or in different accumulations if it is set too low. A good value of winsorization is situational but conventionally winsorization is often at the 90% or 95% quantile, which suggests $d_{\max} = 0.9 \max\{\tilde{r}_i\}$ or $d_{\max} = 0.95 \max\{\tilde{r}_i\}$ for the defined $\tilde{r}_i$.

3.2 The OPTICS Cordillera

We can now turn to give a concrete instance of the Cordillera applied to the pair $(R(X), \{r^*_{(s)}\})$ where $R(X)$ is the OPTICS ordering from Algorithm 1 and $r^*_{(s)} = r^*_{\rho(i)} = r^*_i$ for x_i is the representative reachability as defined in (1). The (absolute) OPTICS Cordillera (OC) is then

$$OC(X; k, \epsilon_{\max}, d_{\max}, q) = \left(\sum_{s=2}^{N} |r^*_{(s)} - r^*_{(s-1)}|^q \right)^{1/q} \tag{2}$$

Loosely speaking, the Cordillera sequentially takes a function of a representative distance between two accumulations and subtracts a function of representative distances within the two accumulations from it, which then gets aggregated; this trades off the accumulation separation with the accumulation density in some way. In the OPTICS Cordillera the representative distance within an accumulation is the minimum representative reachability in an accumulation, the representative distance between two accumulations is their single linkage distance and the points from which the representative within and between distances are calculated are found by the OPTICS algorithm.

The usefulness of winsorization becomes apparent. OC aggregates the absolute representative reachability differences over $R(X)$. If there is a large difference in representative reachability (say, due to outliers) the OC value will be high. If this representative reachability is winsorized to $d_{\max}$, the OC value will be equal or less. Note that $d_{\max}$ may also winsorize the distance between accumulations or even distances within accumulations if it is chosen too small. It should therefore be used sensibly only for making the OC robust against large outlying $r^*_{(s)}$.

Upper and lower bounds for the OPTICS Cordillera. The OPTICS Cordillera in (2) is bounded. The lower bound is 0. A non-trivial upper bound for the observed OPTICS Cordillera in case of maximal clusteredness as a function of N and k can be derived.

Proposition 1 (Bounds of the OPTICS Cordillera.)**.** For $d_{\max} > 0$ we have

$$0 \leq OC\left(X; k, \epsilon_{\max}, d_{\max}, q\right) \leq OC_{\max}\left(X, d_{\max}, k, q; \epsilon_{\max}\right)$$

with

$$OC_{\max}\left(X, d_{\max}, k, q; \epsilon_{\max}\right) = \sqrt[q]{d_{\max}^q \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)}$$

A proof can be found in Appendix A. The bound is sharp for $n = N/k$.

Normalization of the OPTICS Cordillera. We suggest to use Proposition 1 to normalize (2) to lie between 0 and 1. The normalized OPTICS Cordillera, OC', is given by

$$\begin{aligned} OC'(X; k, \epsilon_{\max}, d_{\max}, q) &= \frac{OC(X; k, \epsilon_{\max}, d_{\max}, q)}{OC_{\max}\left(X, d_{\max}, k, q; \epsilon_{\max}\right)} \\ &= \left(\frac{\sum_{s=2}^{N} |r^*_{(s)} - r^*_{(s-1)}|^q}{d_{\max}^q \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)} \right)^{1/q}. \end{aligned} \tag{3}$$

From the denominator expression of (3) the choice of $d_{\max}$ influences the interpretation of OC', see Section 3.2.

Illustration. Figure 3 illustrates the OPTICS Cordillera. In the right column we find the OPTICS Cordillera and the plot of the clusteredness-representative ordering-reachability pair for the representations in the left column. We use $q = 1$ here. The grey barplot shows the $r^*_{\rho(i)}$ on the y-axis for the ordering $R(X)$ of the $x_{\rho^{-1}(\rho(i)}$ on the x-axis. The OPTICS Cordillera is proportional to the length of the black line (displayed up to a constant). The longer this line is, the more clustered the representation is. The Cordillera reaches a minimum if all points have equal $r^*_{(s)}$. The length of the bottom right absolute OPTICS Cordillera is also the upper bound for maximal clusteredness for $N = 4$ and $k = 2$ (with $d_{\max} = \max_g \max_i r_i^{*(g)},\ g = 1, \ldots, 3$).

Properties of the OPTICS Cordillera. The index in (2) has appealing properties for measuring clusteredness.

First, the OPTICS Cordillera can be considered a non-parametric statistic as we define an accumulation solely by the minimum number k of density-connected objects it comprises (Definition 1). This frees the OC from making any stronger assumptions; it inherits from OPTICS the property that the geometrical shape of the accumulations or distribution of objects within the accumulation can be “arbitrary” beyond the effect the used distance measure may have (see Ester et al., 1996, for a discussion what constitutes an arbitrary shape in this setting). Also, nested accumulations of varying density are considered simultaneously (Ankerst et al., 1999).

Additionally, the OC has properties corresponding to the aspects of clusteredness (for fixed meta-parameters) and is therefore suitable to quantify the concept from Section 2. We only paraphrase the properties here; they are formalized and established in the document in the supplementary material.

1. If the distances between the accumulations increase, the OC value does not decrease and typically increases (“Emphasis property”).

2. A denser accumulation of objects around the object with minimal representative

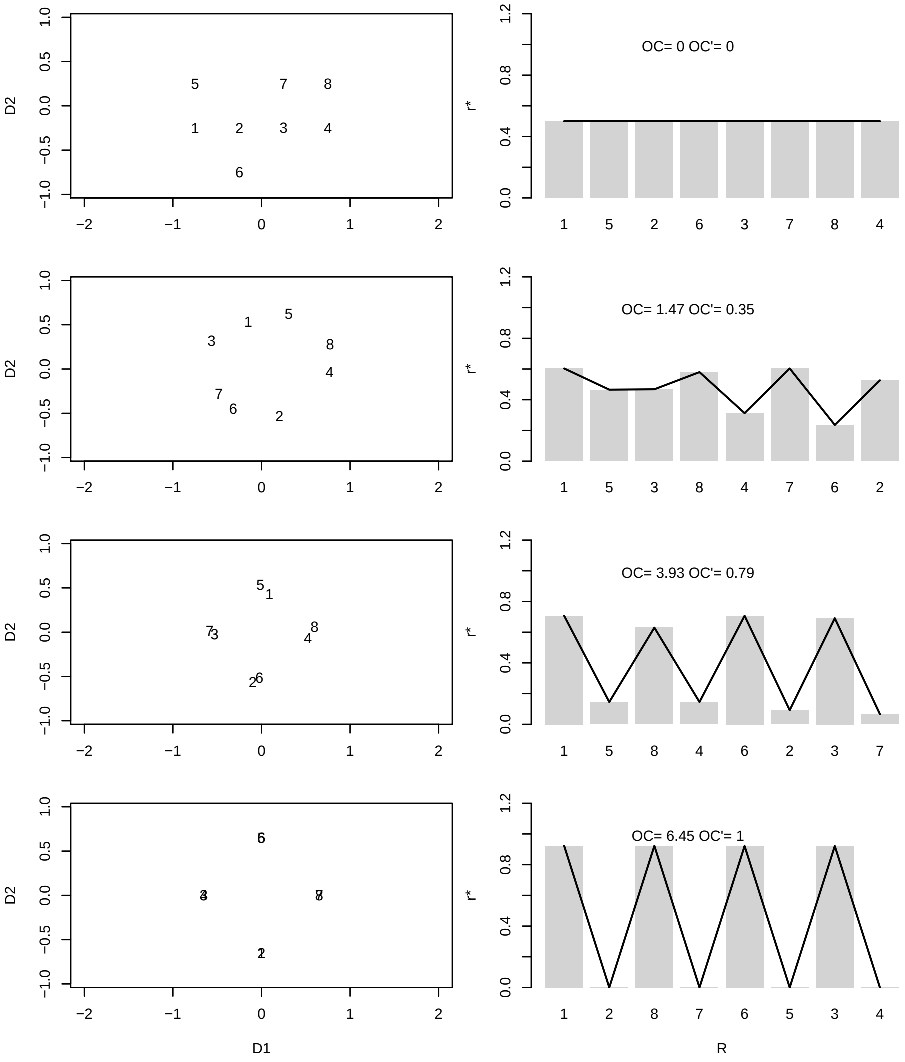


Figure 3: Differently clustered 2D representations of 8 points and their OPTICS Cordillera. In the left column we find $g = 1, \ldots, 4$ representations. The top left plot shows a case of no clusteredness, the bottom left panel shows maximal clusteredness for $N = 8$ and $k = 2$. The other two panels shows representations between these extremes. Clusteredness increases from top to bottom. In the right column we find the corresponding OPTICS reachability plots and with the black line an illustration of the derived clusteredness index, the absolute OPTICS Cordillera (which is here proportional to the real value). The plots are labeled with the numeric value for the normalized OPTICS Cordillera with individual $d_{\max}^{(g)} = \max\{r_{(s)}^{*(g)}\}$. It has been calculated with $k = 2, \epsilon = 2, q = 1$.

reachability will lead to a non-decreasing and typically increasing OC value ("Density property").

3. For an increase in the number of accumulations, the OC value does not decrease and typically increases ("Tally property").

4. For a norm induced metric and sufficiently large $d_{\max}$, if X is radially expanded by a factor $|a| \geq 1$ and the expansion is not offset by change in cluster density, then the OC value does not decrease and typically increases ("Spread property").

It is important to note that these properties usually affect the OC simultaneously and interdependently. Some properties can also work against each other. For example, the spread property and the density property can work against each other as a change in the spread of the points can lead to less dense accumulations and it is conceivable that the decrease in density can offset the change induced by expansion. It is therefore difficult to provide a general characterization of all possible ways the properties can work together in the aggregation.

For local changes in X relating to only one property, however, we can give a rough appraisal: The index can be viewed to comprise two things at once, (i) the aggregation of the differences in representative reachability and (ii) the specific ordering of points by OPTICS. For the effects on (ii) we refer to the original paper Ankerst et al. (1999) and related publications. On the level of the aggregation and for a given ordering, the effect of the properties on the numeric value of OC is governed by q. If $q = 1$ the combination is additive and roughly linear (with an eventual cut-off effect of $\epsilon_{\max}$ or $d_{\max}$). For example (and all else equal), if the distance between two accumulations increases by $|c_1|$ the OC value increases by $|c_1|$. If the smallest representative reachability in an accumulation decreases by $|c_2|$ the OC value increases by $|c_2|$. If both change as described, then the OC value increases by $|c_1| + |c_2|$. If an additional accumulation (with smallest representative reachability in the cluster of c_3) is placed at a distance of c_4 from its closest neighbouring accumulation (with smallest representative reachability in the accumulation of c_5), then the OC value increases by $|(c_4 - c_5)| + |(c_4 - c_3)|$. This assumes the new accumulation forms outside the convex hull of the old accumulations, otherwise it is difficult to characterize the effect generally. If $q > 1$, the representative reachability differences get non-linearly transformed and non-additively aggregated, see (2). On the scale of $OC^q(X)$ the aggregation is then additive for the transformed differences. In general for $q > 1$ the differences are taken to a power and thus larger differences have a higher influence, so more relative emphasis is placed on emphasis, density and spread and relatively lower emphasis on the number of accumulations.

Interpretation and Usage of the OPTICS Cordillera The Cordillera is the $q-$norm in the vector space of differences in subsequent representative reachabilities over the clustering-representative ordering. Interpretation can be guided by the fact that higher values entail any combination of denser, more separated or more accumulations, or more spread out points. Due to the nature of aggregating all of these properties into a unidimensional statistic, the OC does not lend itself to the same detailed interpretation of the overall clusteredness structure as the $N-$dimensional pair $(R(X), \{r^*_{(s)}\})$ does. The OC and ordering-reachabilities pair can be used in combination very much like the average Silhouette (Rousseeuw, 1987) can be used together with the Silhouette plot.

The normalized Cordillera (3) uses as the upper bound the Cordillera value for the most clustered appearance of the N points given $k, d_{\max}$ and all radii up to $\epsilon_{\max}$ to be in the interval $[0, 1]$, with 1 being the most clustered appearance if $N/k \in \mathbb{N}^+$. Accordingly, a normalized OC value can be given the interpretation of a goodness-of-clusteredness statistic, the amount of clusteredness achieved relative to the most clusteredness achievable for a given $d_{\max} = \max_i r^*_i$ of a single representation. It also allows to normalize the index so that one can meaningfully compare a series of representations $X^{(1)}, \dots, X^{(G)}$ with respect to clusteredness if $OC_{\max}(X, d_{\max}, k, q; \epsilon_{\max})$ is constant for all G results, e.g., set $d_{\max}(X^{(1)}, \dots, X^{(G)}) = \max_g \max_i r^{*(g)}_i$ for $g = 1, \dots, G$. In this case the ordering of the Cordillera values entails an ordering of clusteredness. We show examples of these usages in the next section. The third possibility is to set $d_{\max}$ to an *a priori* constant value, e.g., $\epsilon_{\max}$ or some other value above which winsorization takes place; then the interpretations are analogous to the maximum allowed $d_{\max}$. Both the absolute and normalized OC may also be used for automation or as a criterion for optimization.

Another use is to apply the OC for one or more representations over a grid of hyperparameters, particularly k and $d_{\max}$ but also different $\epsilon_{\max}$ (for defining noise regions). This can help gain a rich understanding of the clusteredness of specific data representations for different cluster definitions.[2] We show an example of this in Table 1.

4 Application

To illustrate the usage of *OC* we look at a data set of 58 Californian counties for which we have records on 48 observed and projected indicators for climate change related natural hazards such as county averaged 95th percentile daily maximum temperature, projected average number of days where the daily maximum temperature exceeds the high-heat threshold, percentage of a county's census block area vulnerable to unimpeded coastal flooding, projected annual actual evapotranspiration, projected annual baseflow, projected annual wildfire risk, projected annual fractional moisture in the entire soil column and projected annual precipitation. Projections were made based on the IPCC high emission scenario (A2) and the moderate emission scenario (B1) (Nakicenovic and Swart, 2000) for years from 2000 to 2099 by county. The data were compiled from Cooley et al. (2012), California Energy Commission (2008), Pacific Institute (2009). The data set is available in the supplemental R package.

We subject the data set to six dimension reduction techniques for visualisation and representation in two dimensions: PCA (e.g., Jolliffe, 2002), locally linear embedding (LLE; Roweis and Saul, 2000), Sammon mapping (Sammon, 1969), Isomap (Tenenbaum et al., 2000), Power-Stress MDS (POST-MDS; Buja et al., 2008, Groenen and De Leeuw, 2010) and t-SNE (van der Maaten and Hinton, 2008) and explore the clusteredness structure in these results.

We are particularly interested in clusters of at least three counties. For Isomap and LLE we used 3 as the parameter for the neighbourhoods, POST-MDS was fitted with $\kappa = 1.7, \lambda = 4, \nu = 1$. Perplexity for t-SNE was 3. The points obtained from LLE were slightly jittered for readability.

Figure 4 shows the plots from left to right in descending order based on the OC' value.

[2] This is similar to how the K and L functions (Ripley, 1976) for detecting deviations from spatial homogeneity are used for different radii.

In the bottom row are the corresponding reachability plots (grey bars) with an illustration of the OPTICS Cordillera (black line) as well as the OC and OC' values for $k = 3, \epsilon_{\max} = 10$ and $q = 2$ shown. For all situations $d_{\max} = 1.22$ (the largest representative reachability for the PCA result). The $OC_{\max}$ is 56.559.

The highest clusteredness we find for the LLE result with its four extremely dense, spherical accumulations of at least three objects and one less dense accumulation of three points ($OC = 2.592$, $OC' = 0.345$).

The next clustered result is obtained for t-SNE ($OC = 4.488$, $OC' = 0.236$) with a higher number of appreciable accumulations of at least 3 points. While the number of accumulations is higher, the accumulations are less dense (illustrated by the less deep valleys in the OPTICS reachability plot) than for the LLE result—with $q = 2$ the density and emphasis aspects get higher relative weight in the OC than the tally aspect leading to this ranking. Note that there are some clusters here that are not spherical but linear - the Cordillera picks them up nonetheless.

Isomap leads to the third clustered representation ($OC = 1.202$, $OC' = 0.160$). The Isomap representation shows "bridges" between accumulations and thus lower separation of clusters. This is reflected by the comparatively small peaks in the reachability plot. When the reachability plot is cut at the level of 0.4, OPTICS suggests five clusters, including a half moon shape in the middle right (the first valley in the reachability plot) and the four "arms" (last four valleys). Again, this is picked up in the OC value even though the shapes are different. Also note that there are two clusters nested within the half moon shape for smaller ϵ. The OC also measures these valleys fully, so picks the nesting up as well (i.e., the OC is higher as compared to a single valley at this $\epsilon_{\max}$).

The PCA result is next, being considered less clustered than the Isomap result according to the Cordillera ($OC = 1.177$, $OC' = 0.157$). This is mainly because the accumulations are less dense then in the previous results.

Next is the POST-MDS result ($OC = 1.132$, $OC' = 0.151$), which can largely be explained by less separation between groups of three points, also illustrated by the small peaks in the reachability plot. The POST-MDS result illustrates the effect of winsorizing to $d_{\max}$. In the 2D plot of the POST-MDS we can identify two large outliers at $0.4, -4.5$ (San

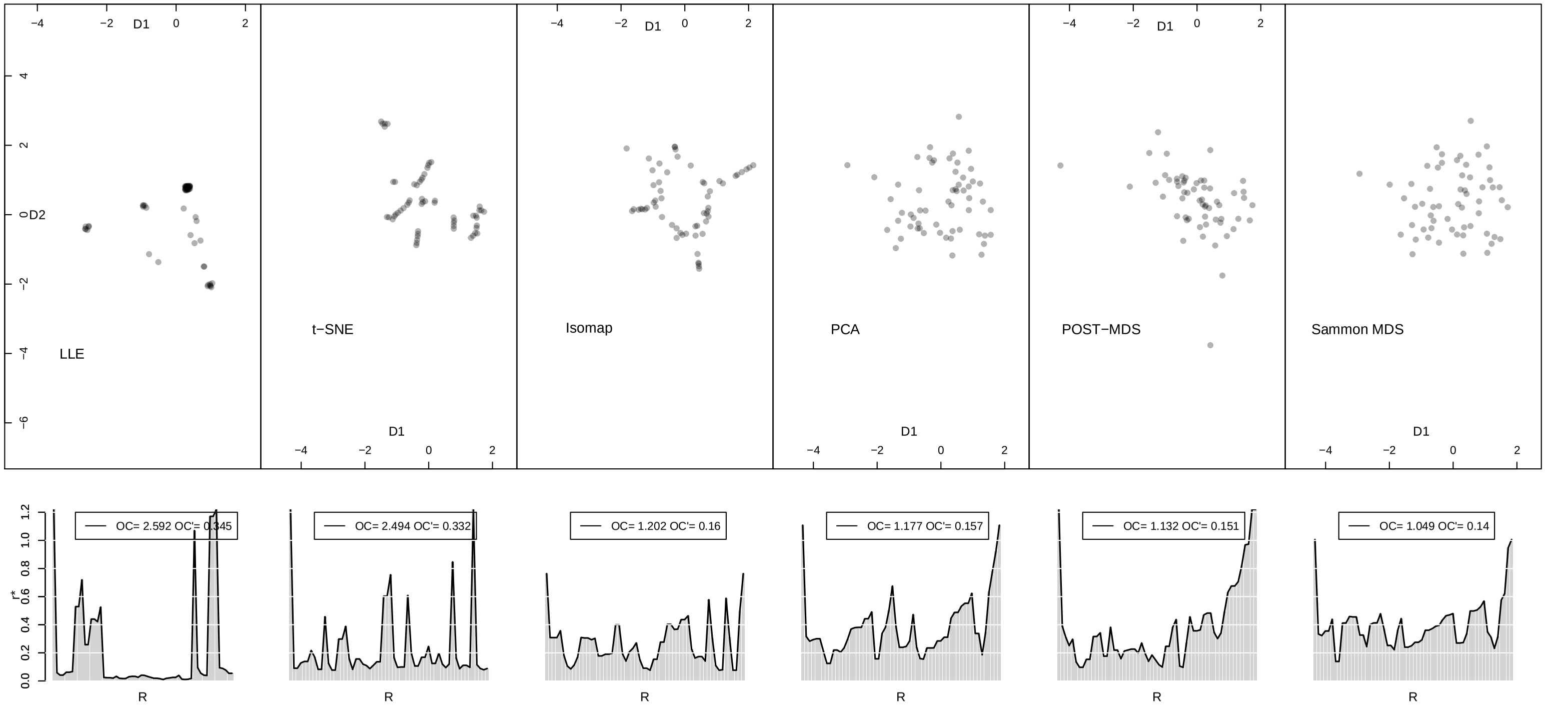


Figure 4: 2D representations of Californian counties based on climate change related natural hazards for six different dimension reduction techniques (top row). From right to left are locally linear embedding (LLE) with $k = 3$, t-SNE, Isomap with $k = 3$, PCA, POST-MDS with $\kappa = 1.5, \lambda = 4, \nu = 1$ and Sammon Mapping. The target dimensions have been scaled to mean=0 and sd=1. The plots are ordered based on the OC' value from left to right, the OPTICS Cordillera values (raw OC and normalized OC') were calculated with $k = 3, q = 2, \epsilon = 10$ and $d_{\max} = 1.22$. The bottom row shows the corresponding OPTICS reachability plots (grey bars) with stylized Cordillera (black).

Francisco) and $-4.3, 1.1$ (Del Norte). With $k = 3$ their suitable reachabilities are > 2, more than twice the largest suitable reachability of all other points 0.97. With setting $d_{\max} = 1.22$ for the OC, we winsorize these two suitable reachabilities to representative reachabilities of 1.22 reducing OC from 5.63 to 1.28. Not winsorizing the suitable reachabilities leads to a higher OC value of the POST-MDS result over, say, the PCA result simply because of these two outliers and their large suitable reachabilities.

The least clustered result is obtained by Sammon mapping ($OC = 1.049$, $OC' = 0.140$). While being similar to the PCA result with respect to accumulation number and accumulation separation, Sammon mapping produces a configuration that shows little density in the accumulations. The core distances with $k = 3$ are rather large as can be seen in the reachability plot and thus the valleys are not very deep. This reduces the OC value.

The ranking obtained by the OPTICS Cordillera is therefore LLE, t-SNE, Isomap, PCA, POST-MDS and Sammon mapping. The two most clustered results are given with county labels in Figure 5.

Similar to the teaser in the motivation section, we asked 37 subjects to rank order the plots in Figure 4 by perceived clusteredness. A parallel coordinates plot of the rankings is given in Figure 6. The human rankings are displayed with grey lines, the ranking obtained by OC with $q = 1$ (dashed) and $q = 2$ (dotted) in black. The ordinate is ordered according to the consensus ranking (maximum $\tau_X = 0.773$; Emond and Mason, 2002) of the human judges.

The rankings are variable but less so then in Figure 1. For $q = 2$ the consensus ranking is largely reflected in the OC—only swapping POST-MDS with PCA. Regarding the latter, POST-MDS had the highest variability in ranks of human judges and when looking at the values of OC', PCA and POST-MDS are numerically rather close. This suggests that for human observers as well as the OC, the two are similarly clustered. One reason for the discrepancy between OC and consensus ranking may be that we used $k = 3$ with the OC, whereas the human judges implicitly used varying $k > 2$. Another may be that $d_{\max}$ is set too low so the higher spread and the outlier in POST-MDS bias its OC value downwards, indicated by PCA having lower OC value than POST-MDS when $d_{\max} > 1.3$.

Note that the judges were divided on whether LLE or t-SNE produces the most clustered

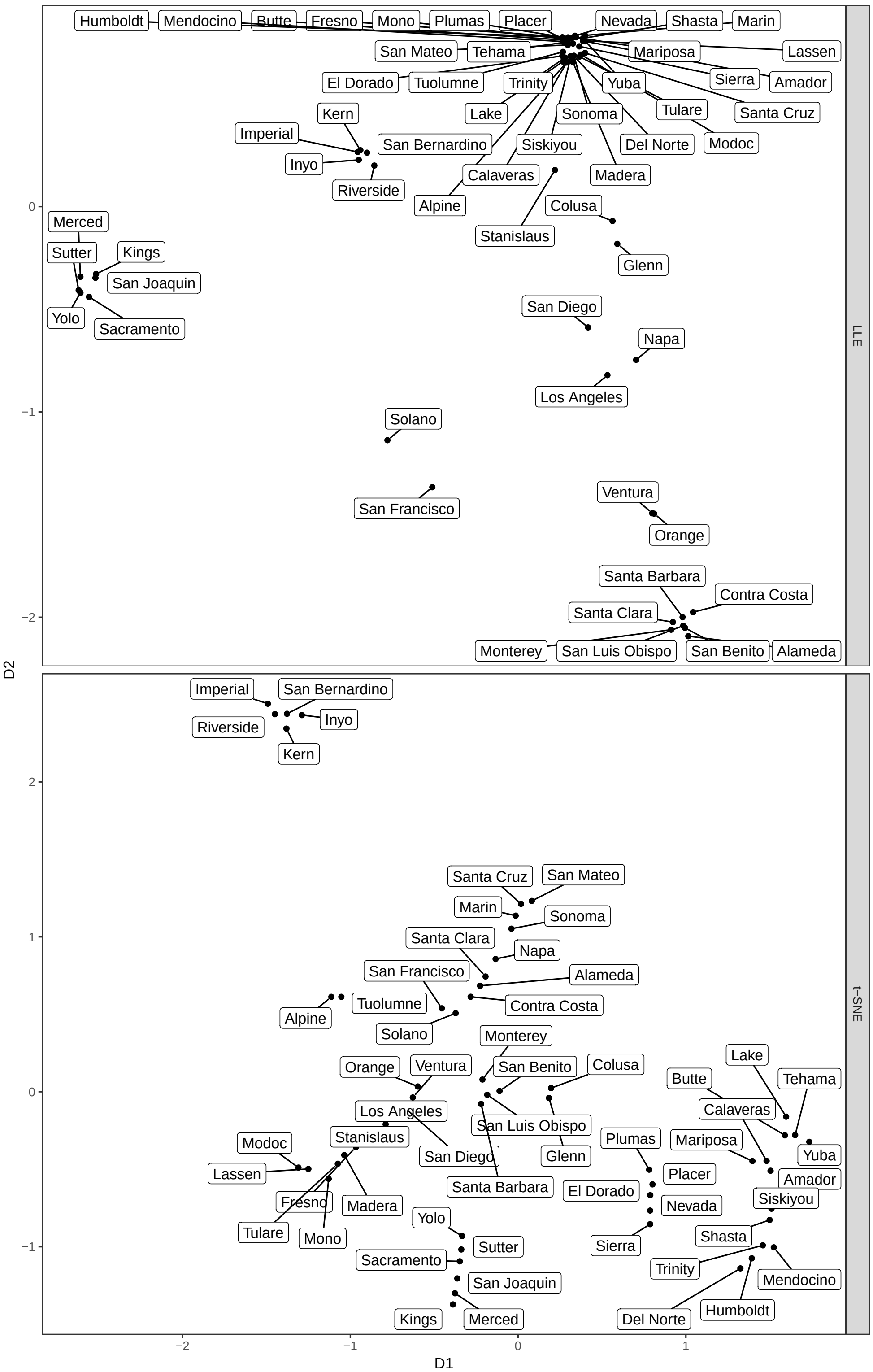


Figure 5: The two most clustered results with county labels.

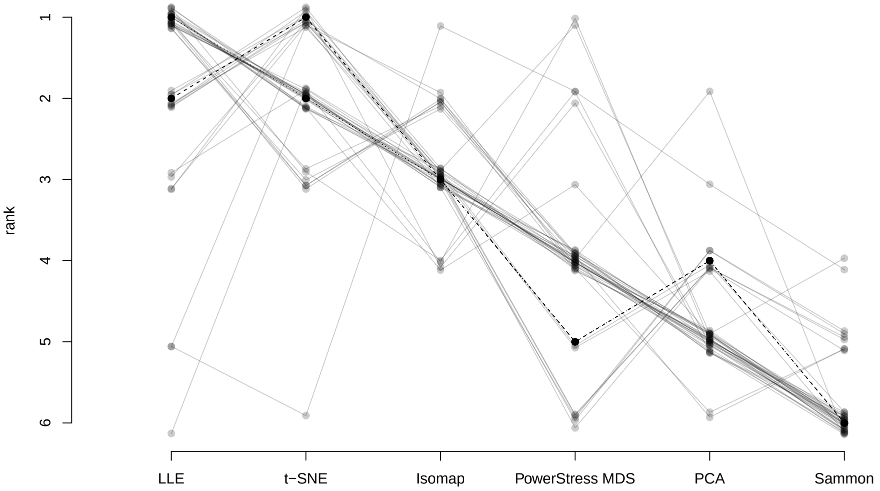


Figure 6: Parallel coordinate plot of rankings (jittered for readability) of 37 subjects of the results in 4 with respect to perceived clusteredness (grey) and the ranking obtained with OC for q=1 (dashed black line) and q=2 (dotted black line). The order on the x-axis reflects the consensus ranking of the human judges.

representation. This can be explained by observing the OC behavior for different q. With $q = 2$, LLE is considered more clustered than t-SNE by OC, but with $q = 1$ they swap places. This can be attributed to a different weighting of the importance of the properties. With $q = 2$ the density and emphasis properties have relatively more weight than the tally property. These two properties have been named the most important aspects by those who chose LLE over t-SNE. With $q = 1$ the larger number of accumulations has more weight—which has been named as an important aspect by most people who chose t-SNE over LLE (but not so much the other way round, see supplementary document). The differences of the rankings obtained by OC when changing q mirrors these two groups of judges.

The choice of meta-parameters can influence the numeric value of OC. This may be used to characterize clusteredness for a grid of hyperparameters. To illustrate these effects, Table 1 lists values for the OC for different parameters $k, \epsilon_{\max}, d_{\max}$ and $q = 1, 2$ for the data representations in Figure 4.

For instance, $\epsilon_{\max}$ is the radius in which to search for neighbours and would be set to a small value only in a very noisy setting (last two rows). When setting it to 0.5 for our examples we naturally find a large reduction in the OC values throughout. In this situation the LLE solution is identified as the most clustered when accumulations must comprise at least $k = 3$ points and POST-MDS when accumulations must comprise at least $k = 10$ points.

For a setting where all points are considered as possible neighbours and we do not treat observations as possible noise ($\epsilon_{\max}$ any large value, here 10), the first four rows of Table 1 list the OC and OC' values for different k and q. We see that LLE produces the most clustered results for $k \geq 3$ and $q = 2$. For $q = 1, k = 3$ t-SNE is more clustered which is due to LLE showing 3-4 very dense accumulations whereas t-SNE shows more accumulations. In general, if q or k is reduced, OC will tend to favor representations that show more accumulations.

The $d_{\max}$ parameter can be used to lessen the effect of outliers and make OC' more robust by winsorizing $r^*_{(s)}$ (rows 6 and 7 in Table 1). POST-MDS produces a number of more outlying points than, for example, the Sammon mapping. When using $d_{\max} = 2.5$ these outliers get more or less full bearing in the normalization and the OC gets larger for POST-MDS. When $d_{\max} = 0.5$ all representative reachabilities even those between accumulations that are larger are cut at 0.5, effectively reducing the relative clusteredness of POST-MDS.

Lastly, one can use the Cordillera as a goodness-of-clusteredness measure relative to the largest representative reachability between any two points of the same representation. In this case an individual $d_{\max}$ for each representation is used, which makes the OC incomparable for different representations. In Table 1 these are the rows 8–9. Here Sammon mapping and POST-MDS are the farthest away from being maximally clustered relative to the largest representative reachability attainable. For $k = 3$ it is t-SNE and for $k = 10$ the POST-MDS result comes closest to the maximal clusteredness possible given the highest observed representative reachability (the latter again due to the outliers).

5 Software

All computing was carried out in R (R Core Team, 2014); the package cordillera accompanying this paper is described in the supplementary material. Further packages used were base (for PCA), lle (Diedrich and Abel, 2012, for LLE), stops (Rusch et al., 2015, for POST-MDS), MASS (Venables and Ripley, 2002, for Sammon Mapping), vegan (Oksanen et al., 2016, for Isomap) and tsne (Donaldson, 2016, for t-SNE). The plots were created by base graphics or with ggplot2 (Wickham, 2009) in combination with ggrepel (Slowikowski, 2016) and plotrix (Lemon, 2006) for the ladderplots.

6 Conclusion and Discussion

For representations of data matrices the question of whether and how clusters of observations form and how well these clusters of observations are visible is often of high interest in data analysis. To be able to do this demands that the appearance is somehow clustered, i.e., there are some appreciable accumulations of observations. In low dimensions this is often assessed by visual interpretation. We observed that the judgement of what makes such a result appear clustered hinges on implicit assumptions which can be very different for different observers. Therefore, the assessment of the clusteredness ultimately lies in

Parameters				PCA		Sammon		LLE		Isomap		Power Stress MDS		t-SNE	
k	$d_{\max}$	q	ϵ	raw	normed	raw	normed	raw	normed	raw	normed	raw	normed	raw	normed
3	1.22	1	10	5.082	0.11	4.409	0.095	7.723	0.167	5.413	0.117	4.849	0.105	9.505	0.205
5	1.22	2	10	0.953	0.163	0.715	0.122	2.481	0.424	0.797	0.136	0.782	0.134	2.087	0.357
10	1.22	2	10	0.638	0.158	0.499	0.123	1.255	0.31	0.428	0.106	0.634	0.157	0.842	0.208
25	1.22	2	10	0.069	0.025	0.08	0.029	1.227	0.45	0.062	0.023	0.383	0.14	0.298	0.109
10	1.22	1	10	2.05	0.153	1.335	0.099	3.363	0.251	1.471	0.11	1.835	0.137	2.426	0.181
3	2.5	2	10	1.177	0.076	1.049	0.068	3.048	0.198	1.202	0.078	2.372	0.154	3.262	0.212
3	0.5	2	10	0.704	0.228	0.71	0.23	1.218	0.395	0.991	0.322	0.699	0.227	1.297	0.421
3	indiv.	2	10	1.177	0.172	1.049	0.169	3.048	0.299	1.202	0.255	2.372	0.169	3.262	0.302
10	indiv.	2	10	1.198	0.212	0.896	0.176	1.43	0.252	0.428	0.128	2.013	0.267	1.584	0.26
3	1.22	2	0.5	0.743	0.099	0.711	0.094	1.109	0.147	0.962	0.128	0.697	0.093	1.255	0.167
10	1.22	2	0.5	0.029	0.007	0	0	0.012	0.003	0.077	0.019	0.132	0.033	0.095	0.024

Table 1: Different values of the OPTICS Cordillera for a grid of different hyperparameter setups for the representations from Figure 4.

the eyes of the beholder. If the representation is given for a higher number of dimensions, the possibility of visualization is severely limited and judging clusteredness is even more difficult.

To make the assessment of a clustered appearance more transparent and reproducible, in this paper we introduced and defined clusteredness in a distance-density based framework as a continuum of appearances between no clusteredness and maximal clusteredness, characterized by a number of aspects used to assess how clustered such results appear including that clusteredness increases when the objects accumulate closer together, the distances between accumulations increases and the number of accumulations increases for a given minimum number of objects in a cluster.

For this operational definition of clusteredness we suggested an index that quantifies clusteredness. This index, the OPTICS Cordillera, is appealing for measuring clusteredness in data representations within a density-distance based framework. It makes weak assumptions on the nature of a cluster including no assumptions on cluster number, cluster shapes, cluster centroids and does not rely on a cluster assignment of observations. Furthermore, the index adheres to the aspects of clusteredness. The index is parsimonious with the number of mandatory parameters but also includes optional parameters that allow to tune the index to different needs including making the index robust to noise points and outliers or weighting the aspects of clusteredness differently. We derived bounds for the index and use them to normalize the index.

For a single data representation, the index can be used as a descriptive goodness-of-clusteredness statistic, e.g., to assess and quantify how close the result is to displaying no clusteredness or maximal clusteredness, or to assess the change of clusteredness relative to different cluster sizes or cluster density specifications. For a series of representations, the index can be used to compare them with respect to their clustered appearance. For a grid of hyperparameters the index can be used to characterize a data representation with respect to clusteredness as a function of minimum number of points in a cluster, different neighbourhood radii or maximum distances. The OPTICS Cordillera may also be used in augmenting loss functions for different methods that inherently need or produce a clustering or classification structure or for hyperparameter selection.

We note that the Cordillera measure and the presented ideas can be extended beyond the ordering obtained with OPTICS, say, to an ordering derived from a minimum spanning tree (Forina et al., 2004). The Cordillera would then simply be measuring the length over another clusteredness-representative ordering-distances pair and allow to capture clusteredness in a different clusteredness framework analogously.

The OC also has its limitations. It was developed for use in exploratory settings and in conjunction with hierarchical, unsupervised procedures. Particularly with partitional clustering, when a decision on what the actual clusters are has to be made or when cluster labels are available, internal cluster validity indices are usually more appropriate. We see our index only as complementary in this case; it may be used to gauge prior to running a clustering algorithm if a density based clustering will lead to a useful result. Furthermore the OC is meant to be interpreted and tuned relative to the situation at hand.

The OPTICS Cordillera is a versatile, flexible index to gauge the structure of clusteredness which might be of interest in many contexts where the tendency of point vectors to accumulate in some space should be assessed. Such cases could be astronomy, where one would want to assess the arrangement of stars in galaxies, or in neuroscience, where it might be of interest to find out how the activation pattern of neurons in a brain is represented for different tasks.

A Appendix A: Proof

Proof of Proposition 1 (Bounds for the OPTICS Cordillera). Proposition 1 can be shown by establishing the upper and lower bound of $OC(X; \epsilon_{\max}, d_{\max}, k, q)$ for a given $\epsilon_{\max}, d_{\max}k, q$, so $OC(X)$.

For the lower bound observe that $OC(X) \geq 0$ as $OC(X)$ is a sum of non-negative, $|r^*_{\rho(i)} - r^*_{\rho(i)-1}|$ so the left hand side in Proposition 1 follows. □

For the upper bound, we first look at the $q-$th power of $OC(X)$, $OC(X)^q$ which is additive in $|r^*_{\rho(i)} - r^*_{\rho(i)-1}|^q$. We use the arrangement described by Definition 3. From the OPTICS algorithm there is a distinct seesaw pattern connected with Definition 3 after applying Algorithm 1. The result is the pair $(R(X), \{r^*_{\rho(i)}\})$ stemming from the deterministic Algorithm 1 which can be characterized in the following way (if Definition 3 applies):

For points x_i, x_j in the same accumulation, $r_{ij} = c_i = c_j = 0$. When the x_j get ordered sequentially as in OPTICS, there are for all $x_j \in C(x_i)$ then $(k-1) \times \tilde{r}_{\rho(j)} = r^*_{\rho(j)} = 0$ as representative reachabilities of $x_{\rho(j)}$. After all points in an accumulation have been processed, the algorithm turns to a point from the closest neighbouring accumulation. For points x_i and x_l in neighbouring accumulations $r_{il} = d_{\max}$ and thus $r^*_{\rho(l)} = d_{\max}$. Since all points in a cluster again get processed sequentially by OPTICS the whole process repeats. The first single linkage reachability $\tilde{r}_{(1)}$ is undefined, so we set $r^*_{(1)} = d_{\max}$, see (1). The ordering thus consists of a repeating pattern of $r^*_{(s)} = d_{\max}$ signalling the beginning of an accumulation, followed by $(k-1) \times 0$, so $r^*_{(s+1)}, ..., r^*_{(s+l-1)} = 0$ for points $x_{\rho^{-1}(s)}, \ldots, x_{\rho^{-1}(s+l-1)}$ in the same accumulation; this then gets repeated with the closest neighbouring accumulation, so we have $r^*_{(s+l)} = d_{\max}$ for $x_{\rho^{-1}(s+l)}$ in the closest neighbouring cluster and then again $r^*_{(s+l+1)}, ..., r^*_{(s+2l-1)} = 0$. This pattern of one $d_{\max}$ and $(k-1) \times 0$ repeats as often as there can accumulations of size k be formed.

Note that this pattern is direct consequence from applying OPTICS to Definition 3 and is unique only up to permutations of accumulations in $R(X)$ or permutations of points for a given accumulation in $R(X)$. The OC however is invariant to these permutations.

For the differences of $|r^*_{(s)} - r^*_{(s-1)}|$ to be maximal we must have either $|d_{\max} - 0|$ or $|0 - d_{\max}|$. Under Definition 3 this can happen only for observations between accumulations; within an accumulation this difference is 0. We thus need to count the maximum possible number, o, of accumulations of $k-1$ points with $r^*_{\rho(j)} = 0$ as for each of these accumulations there must be at most two jumps from and to an observation x_l with $r^*_{\rho(l)} > 0$. Because of the additivity of the elements of $OC(X)^q$, in the maximally clustered case this must satisfy

$$\begin{aligned} N &\leq o(k-1) + t \\ o &\leq t \leq o + 1 \end{aligned}$$

with t being the number of points with $r^*_{\rho(l)} > 0$. Substituting the second equality into the first leads after algebraic manipulation to

$$\frac{N-1}{k} \leq o$$

If OPTICS cannot order the points for these identity to hold exactly, then the above identity is an upper bound. Since o must be integer this means the next closest o fulfilling this is

$$o = \left\lceil \frac{N-1}{k} \right\rceil$$

This means the number of jumps in the reachability plot from a group of observations with $r^*_{\rho(j)} = 0$ to $r^*_{\rho(l)} > 0$ or back is at most

$$2 \left\lceil \frac{N-1}{k} \right\rceil$$

and since the maximum possible length of the jump is $d^q_{\max}$, with maximal clusteredness we have

$$OC(X)^q \leq d^q_{\max} 2 \left\lceil \frac{N-1}{k} \right\rceil$$

This bound can be improved for the case where the last group has no last jump anymore by subtracting a single $d^q_{\max}$. This means therefore

$$OC(X)^q \leq d^q_{\max} \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)$$

Taking the $q-$th principal root of the above identity leads to

$$OC(X) \leq \sqrt[q]{d^q_{\max} \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)} \quad \square$$

B Appendix B: The OPTICS Algorithm

OPTICS defines two distances:

Definition 5 (Core Distance)**.** The core distance c_i is the distance of a vector x_i to the $k-$th closest points

$$c_i = c(x_i; k, \epsilon_{max}) = \begin{cases} \max(d_{ij}) : j \in S_{k,\epsilon_{\max}}(x_i) & \text{if } S_{k,\epsilon_{\max}}(x_i) \neq \emptyset \\ \text{undefined} & \text{if card}\left(N_{\epsilon_{\max}}(x_i)\right) < k \end{cases} \quad (4)$$

Definition 6 (Reachability Distance)**.** The reachability distance r_{ij} between two points x_i and x_j is the maximum of d_{ij} or c_i, so

$$r_{ij} = r(x_i, x_j; k, \epsilon_{\max}) = \begin{cases} \max\left(c_i, d_{ij}\right) & \text{if } S_{k,\epsilon_{\max}}(x_i) \neq \emptyset \\ \text{undefined} & \text{if card}\left(N_{\epsilon_{\max}}(x_i)\right) < k \end{cases} \quad (5)$$

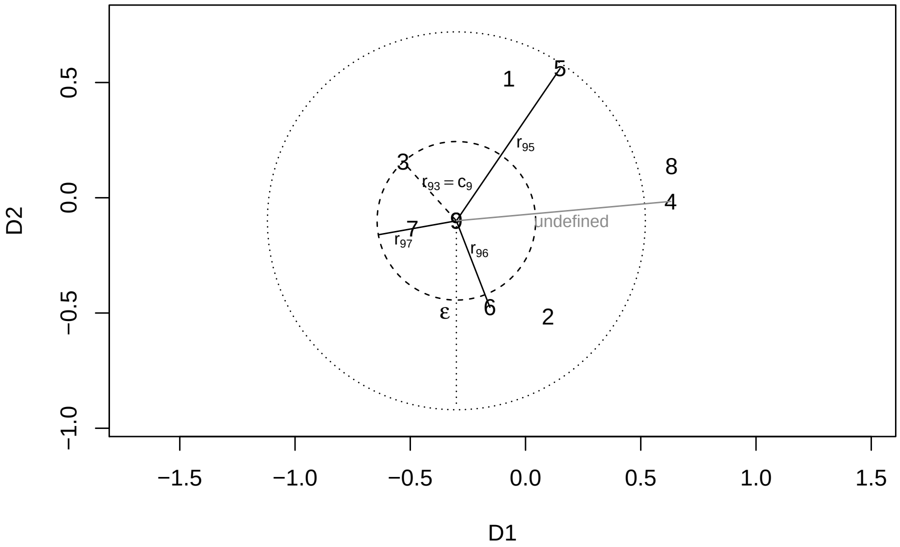


Figure 7: Two-dimensional illustration of the distances c_i and r_{ij} used in OPTICS . The parameters are $k = 3$ and $\epsilon = 0.82$. We use point x_9 as the reference. The region around x_9 in which to look for neighbours is defined by ϵ_{max} and is illustrated by the dotted circle. The core distance of x_9, c_9 is the distance to the $k-$th closest point including x_9 (x_3). The core distance is roughly 0.34 (dashed line). The core region around x_9 is illustrated by the dashed circle. The reachability distance between x_9 and any other point x_j, r_{9j}, is the maximum of core distance or direct distance or is undefined if the point falls beyond the ϵ_{max} radius. The illustrated as solid lines for a few examples, $r_{93} = c_9$, $r_{97} = c_9$, $r_{96} = d_{96}$, $r_{95} = d_{95}$ and $r_{94} =$ undefined.

A graphical representation of these distances is shown in Figure 7. The parameters are $k = 3$ and $\epsilon_{\max} = 0.82$, x_9 is the reference point. The region around x_9 in which to look for neighbours is defined by $\epsilon_{\max}$, the dotted circle. The core distance of x_9, c_9, is the distance to the $k-$th closest point (including x_9) which is point x_3. The core distance is roughly 0.34, the length of the dashed line. It is also r_9^*. At $k = 3$ and $\epsilon_{\max} \geq c_9$ all points including x_9 within the distance c_9 around x_9 are core points and are directly density reachable from x_9. The core region around x_9 is illustrated by the dashed circle. The set of these points is $S_{3,\epsilon_{\max}}(x_9) = \{9, 7, 3\}$ and the core distance is the maximum distance to any of the points in $S_{3,\epsilon_{\max}}(x_9)$.

The reachability distance between x_9 and any other point x_j, r_{9j}, is the maximum of core distance of x_9 or direct distance of d_{9j} or is undefined if the point falls beyond the $\epsilon_{\max}$ radius. This is illustrated as the length of the solid lines for a few examples. For example for r_{97} it is $\max(c_9, d_{97})$ which is c_9, for r_{96} it is $\max(c_9, d_{96})$ which is d_{96} and for r_{94} it is undefined as x_4 is more than $\epsilon_{\max} = 0.8$ distant from x_9. This illustrates the function and optionality of $\epsilon_{\max}$: Any $\epsilon_{\max}$ will contain the defined distances of any smaller ϵ (i.e., denser accumulation) which enables the simultaneous characterization of many accumulations with different densities between objects up to $\epsilon_{\max}$. Thus, $\epsilon_{\max}$ needs not necessarily be set but can be just large. Setting $\epsilon_{\max}$ will lead to treating points further away as noise instead of a neighbour (here, x_8 and x_4).

The algorithm is then:

Algorithm 1 A pseudo code representation of the main OPTICS algorithm (upper part) and the update function (after Ankerst et al. (1999) and Wikipedia (2015)).

```
OPTICS(Data, epsilon, k)
    empty ordered list
    FOR i FROM 1 to N of Data
       x=x_i
    IF (processed(x) == FALSE)
       S = neighbors(x, epsilon)
       set x as processed
       x.reachability-distance = UNDEFINED
       x.core-distance = core-distance(S,epsilon,k)
       output x to ordered list
       IF (x.core-distance != UNDEFINED)
          OrderSeeds = empty priority queue
          update(OrderSeeds, S, x)
          WHILE (empty(OrderSeeds)==FALSE) DO
             y = next(OrderSeeds)
             S'= neighbors(y, epsilon)
             set y as processed
             y.core-distance = core-distance(S',epsilon,k)
             output y to the ordered list
             IF (core-distance(y, epsilon, k) != UNDEFINED)
                update(OrderSeeds, S',y)
END

update(OrderSeeds, S, x)
    coredist = x.core-distance
    FOR EACH y IN S
       IF (processed(y) == FALSE)
          new-reach-dist = max(coredist, distance(x,y))
          IF (y.reachability-distance == UNDEFINED)
              y.reachability-distance = new-reach-dist   //y not in OrderSeeds
              insert(OrderSeeds, y, new-reach-dist)
          ELSE               // y is in OrderSeeds, check for improvement
              IF (new-reach-dist < y.reachability-distance)
                 y.reachability-distance = new-reach-dist
              moveup(OrderSeeds, y, new-reach-dist)
END
```

SUPPLEMENTARY MATERIAL

Supplementary Document: A supplement with the results from the qualitative study on clusteredness perception and with details and proofs of the clusteredness properties of the OPTICS Cordillera. (`cordillera-supplement.pdf`, PDF file)

R Package: R-package `cordillera` containing an implementation of the OPTICS Cordillera described in the article. The package also contains all data sets used as examples in the article. (`cordillera_0.6-0.tar.gz`, GNU zipped tar file)

R Script: A file to reproduce the results, tables and figures of the paper. (`cordillera-script.R`, text file)

README: A README file. (`README`, text file)

All supplemental files are contained in a single archive. (`cordillera-supplement.zip`, ZIP file)

References

Alimoglu, F. (1996). Combining multiple classifiers for pen-based handwritten digit recognition. Master's thesis, Bogazici University, Istanbul, Turkey.

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander (1999). OPTICS: Ordering points to identify the clustering structure. In *ACM SIGMOD International Conference on Management of Data*, Volume 28, pp. 49–60. ACM Press.

Buja, A., D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics 17*(2), 444–472.

California Energy Commission (2008). Raster downloads. [accessed July 14, 2014].

Cooley, H., E. Moore, M. Heberger, and L. Allen (2012, July). Social vulnerability to climate change in California. Technical Report Publication Number: CEC-500-2012-013, Pacific Institute, California Energy Commission. [accessed, July 16, 2014].

Diedrich, H. and M. Abel (2012). *lle: Locally linear embedding.* R package version 1.1.

Donaldson, J. (2016). *tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE).* R package version 0.1-3.

Emond, E. J. and D. W. Mason (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis 11*(1), 17–28.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231. AAAI Press.

Forina, M., M. C. Oliveros, C. Casolino, and M. Casale (2004). Minimum spanning tree: ordering edges to identify clustering structure. *Analytica Chimica Acta 515*(1), 43 – 53.

Greenacre, M. (2011). A simple permutation test for clusteredness. Technical Report 555, University Pompeu Fabra, Barcelona, Spain.

Groenen, P. and J. De Leeuw (2010). Power-stress for multidimensional scaling. Technical report, UCLA, Los Angeles, USA.

Huang, H., Y. Liu, M. Yuan, and J. S. Marron (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics 24*(4), 975–993.

Jain, A. K. and R. J. Dubes (1988). *Algorithms for clustering data.* Englewood Cliffs: NJ: Prentice Hall.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News 6*(4), 8–12.

Liu, Y., Z. Li, H. Xiong, X. Gao, and J. Wu (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 911–916. IEEE.

Nakicenovic, N. and R. Swart (Eds.) (2000). *Emission scenarios.* Cambridge, UK: Cambridge University Press.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner (2016). *vegan: Community Ecology Package.* R package version 2.4-1.

Pacific Institute (2009). Census blocks, percent flooded under sea level rise scenarios [csv data file]. [accessed July 9, 2014].

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org/.*

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability 13*(2), 255–266.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*(5500), 2323–2326.

Rusch, T., J. De Leeuw, and P. Mair (2015). *stops: STructure Optimized Proximity Scaling.* R package version 0.0-17, Available at *http://r-forge.r-project.org/projects/stops/.*

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*(5), 401–409.

Slowikowski, K. (2016). *ggrepel: Repulsive Text and Label Geoms for 'ggplot2'.* R package version 0.6.5.

Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319–2323.

Tibshirani, R. and G. Walther (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics 14*(3), 511–528.

van der Maaten, L. and G. Hinton (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9*(Nov), 2579–2605.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Wang, X., W. Qiu, and R. H. Zamar (2007). Clues: A non-parametric clustering method based on local shrinking. *Computational Statistics & Data Analysis 52*(1), 286 – 298.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wikipedia (2015). Optics algorithm — Wikipedia, the free encyclopedia. [Online; accessed 10-October-2015].