Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons

Thomas Lefèvre, Patrick Chariot, Pierre Chauvin

Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons.

Thomas Lefèvre[1,2,3,4], Patrick Chariot[3,4], Pierre Chauvin[1,2]

[1]Inserm, UMRS 1136, Pierre Louis Institute of Epidemiology and Public Health, Department of Social Epidemiology, Paris, France

[2]Université Pierre et Marie Curie-Paris 6, UMRS 1136, Paris, France

[3] AP-HP, Hôpital Jean-Verdier, Department of Forensic Medicine, F-93140 Bondy, France

[4] IRIS - Institut de recherches interdisciplinaires sur les enjeux sociaux (INSERM, CNRS, EHESS, Université Paris 13, UMR 8156-723), Bobigny, France

Phone: +33 (0)144738460

Fax: +33 (0)144738663

Institutional address: Inserm, UMRS 1136, 27 rue Chaligny, 75571. Paris Cedex 12, France

E-mail addresses: thomas.lefevre@inserm.fr, patrick.chariot@jvr.aphp.fr, pierre.chauvin@inserm.fr

Corresponding author: Thomas Lefèvre

ABSTRACT

Researchers handle increasingly higher dimensional datasets, with many variables to explore. Such datasets pose several problems, since they are difficult to handle and present unexpected features. As dimensionality increases, classical statistical analysis becomes inoperative. Variables can present redundancy, and the reduction of dataset dimensionality to its lowest possible value is often needed. Principal components analysis (PCA) has proven useful to reduce dimensionality but present several shortcomings. As others, forensic sciences will face the issues specific related to an evergrowing quantity of data to be integrated. Age estimation in living persons, an unsolved problem so far, could benefit from the integration of various sources of data, e.g. clinical, dental and radiological data.
We present here novel multivariate techniques (nonlinear dimensionality reduction techniques, NLDR), applied to a theoretical example. Results were compared to those of PCA. NLDR techniques were then applied to clinical, dental and radiological data (13 variables) used for age estimation. The correlation dimension of these data was estimated.
NLDR techniques outperformed PCA results. They showed that two living persons sharing similar characteristics may present rather different estimated ages. Moreover, data presented a very high informational redundancy, i.e. a correlation dimension of 2.
NLDR techniques should be used with or preferred to PCA techniques to analyze complex and big data. Data routinely used for age estimation may not be considered suitable for this purpose. How integrating other data or approaches could improve age estimation in living persons is still uncertain.

Keywords: Nonlinear dimensionality reduction; clustering; age estimation; multivariate methods; big data

1. Introduction

Scientific research and forensic science are – at least partly – about finding associations between factors and searching for underlying causal relationships between so-called exposure and events of interest. Whether accounting for multiple covariates to control for possible biases or not, classical hypothesis testing and linear regression are extensively used and have proved to be relevant. However, they may not be sufficient to address all issues and analytical needs in forensic sciences [1,2]. Widening the scope of experimentation in forensic sciences, a key question can be raised: how close is one subgroup of people to another or how different are they? Clustering techniques are used in different fields, such as computer vision and imaging [3], genetics [4] and public health [5]. The international ENCODE project made extensive use of clustering techniques to systematically search for the functionality of “junk” DNA [6]. These methods provide clues for identifying homogeneous groups of people, but they share a common limitation: all of them operate on a “flat” feature space. In the real world, the neighborhood or proximity between two persons may not respect such a geometric assumption and inappropriate shortcuts may appear, falsely linking two people who should not be related to each other. Moreover, researchers have to handle increasingly large datasets, with dozens of potential outcomes and as many explanatory variables of interest. Classical clustering methods and classical statistical tools lose their ability to separate two distinct groups as dimensionality increases, as well as their ability to reach statistical significance. This is known as the “curse of dimensionality” or the “empty phenomenon” [7,8]. According to Lee and Verleysen [8], issues with heterogeneous data can appear as soon as we deal with 10 to 20 or more variables. Data should then be considered as “big data” in many cases.

The methods discussed in this paper address two related issues: respecting the intrinsic geometry of data and reducing their dimensionality to make them more comprehensive and even graphically displayable. These methods are called “nonlinear dimensionality reduction” (NLDR) techniques and appeared at the beginning of the 2000s [9,10]. Since then, variants

have been proposed [10,11]. In this article, we first presented the importance of preserving the intrinsic geometry of data. Second, we provided a brief description of a typical NLDR technique and compared its performance with principal components analysis (PCA) and multidimensional scaling (MDS) performances on a theoretical example. Third, we applied NLDR techniques to age estimation in living persons. Estimating the age of a person based on various clinical or non clinical data is an old challenging problem in forensic sciences that is standing still, regardless how often it has been explored [12-16]. Even if the way forensic physicians deal with this topic can be controversial, most experts agree that combining any potential informative data is the best mean to reach accuracy, e.g., combining dental and other radiological data [17-19]. The demonstration of a significant linear trend between different characteristics of a living person and its chronological age is not dubious, but is also poorly contributory to an accurate estimate of the person's age. We actually have no precise idea of how informative are the data usually handled to estimate the age of a living person, and how relevant they are to discriminate a person from another one in terms of age, not to mention to determine if a person reached the legal majority or not. To date, a single study used PCA to estimate the age-at-death [20] and none to estimate the age in living persons. Here, we applied NLDR techniques to empirical forensic data, integrating clinical, dental and radiological data and investigated whether these data could properly and accurately ground age estimation in living people.

2. Methods

2.1 Preserving data geometry and complexity

The difference between a flat space and a more generic, curved space can be likened to the difference between considering the earth to be flat and considering it to be spherical. There is no universal projection method for creating a flat map of the earth with virtually no geometrical distortion, either angular or metric [21]. In the same way, in the absence of information about how datasets are "physically" structured, it may be inaccurate to assume that they are flat, with no curvature at all. Figure 1 shows how geometry determines whether two data points are neighbors or not. It also emphasizes why the dimensionality of a dataset can be reduced.

2.2 Nonlinear dimensionality reduction techniques versus classical techniques

PCA methods can be used to reduce a dataset dimensionality [10,22] by rearranging the feature space by combining variables into factors. These factors are obtained so that they are not correlated with one another. While PCA-like techniques are effective in many cases, their main limitation lies in their linear assumption. The linearity hypothesis assumes that the entire problem to be addressed can be broken down into elementary sub-problems to which the correct weightings can be added to reconstitute the entire initial problem. PCA-like techniques should not provide fundamentally wrong results, although they may destroy evidence for truth or distort more subtle relationships. Therefore, there is room for more suitable techniques than PCA-like techniques. In contrast to the linear assumption in PCA, these techniques are called "nonlinear dimensionality reduction" (NLDR) techniques. They only consider geometrical proximity, apart from statistical considerations. The objective of NLDR techniques is to build the most "respectful" space in terms of "true" neighborhood. For this, these techniques depend on the construction of geodesic paths. NLDR techniques can preserve the nonlinear associations between variables. Data dimensionality is reduced, while the potential complexity of the associations between the variables is qualitatively preserved. The first NLDR algorithms appeared in the early 2000s, the archetype being ISOMAP [9,10]. Other methods have been proposed, based on the construction of a graph depicting the neighbor relationships for each data point. Given an intrinsic dimension and a dataset as inputs, the ISOMAP algorithm operates in three steps, as given in Table 1. Another class of NLDR techniques preserves the topological complexity and properties of datasets (i.e. their neighboring relationships: two individuals close to each other in the initial dataset remain close to each other in the reduced dataset), based on neural networks, such as the autoencoders [23-25].

2.3 Limitations and comparisons of three algorithms: an empirical approach

We ran the ISOMAP technique on a theoretical example. We also included PCA and MDS techniques for comparison. Results are provided by a MATLAB program called MANI (see supplementary material). The present theoretical example is usually called the Swiss roll dataset [9]. The Swiss roll is a plane that is rolled up with none of its surfaces touching any other. Its intrinsic dimension is 2 (data points all belong to a 2-dimensional plane). Using dimensionality reduction algorithms, we plan to unfold them to obtain a rectangular plane for the Swiss roll. The execution times for the different algorithms are indicated into brackets. Performance of other NLDR techniques both applied on the Swiss roll dataset as well as applied on two other theoretical datasets can be found in additional data.

2.4 Parameters tuning

Most NLDR techniques require parameter setting. Apart from the estimated intrinsic dimension, there is usually only one tuning parameter to be inputted. ISOMAP requires defining the neighborhood, that is, how many neighbors should be searched for around a data point. If the data are relatively sparse, specifying a high number of neighbors may lead to register some that are actually far from that point. In other words, it can lead the algorithm leaping undesirably from one surface to the next if they are close to each other and if there are not enough neighbor data points.

2.5 Dimensionality reduction with intrinsic geometry preservation

Before reducing the initial dimensionality of a dataset to its intrinsic one, this intrinsic dimension needs to be estimated without any prior information. The intrinsic dimension of a dataset can be defined as the minimum number of independent variables needed to describe it without information loss [26]. The approaches to retrieving this number from a dataset are based on different conceptions of dimensionality. Camastra proposed the following taxonomy for these different methods: local, global and fractal-based [26]. Several techniques exist for estimating the intrinsic dimension of a dataset. Reviews of these techniques can be found in [26,27]. We give here the example of the correlation dimension, which is also used in system dynamics [28]. It is based on the correlation integral $C_m(r)$, which is defined as:

$$C_m(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(|Xj - Xi| =< r) \qquad \text{(Eq. 1)}$$

where X_i are data points, N the number of data points in the sample, r an arbitrary radius, and I the indicator function (i.e., I(condition) = 1 if the condition is true, 0 if it is not).

The correlation dimension D is then defined as

$$D = \lim_{r\to 0} \frac{\ln(C_m(r))}{\ln(r)} \qquad \text{(Eq. 2)}$$

Techniques other than the correlation dimension exist, such as the nearest-neighbor estimator [26] or the maximum-likelihood estimator [29]. The results of these types of dimensionality estimators applied on four examples of theoretical datasets are reported in Table 2. For each example, we give the intrinsic dimensionality, which is the dimensionality to be retrieved.

2.6 About outliers

Some NLDR techniques are not data-conservative. If an initial dataset consists of 3,000 observations, the processed dataset may contain only 2,500 of them (ISOMAP behaves like this). The reason lies in the graph construction, where distances are estimated and k neighbors are searched for. Some points may appear to be not connected to any others and are assigned an “infinite” distance to the other points. In such a case, they are isolated and eliminated from the dataset since they are seen as outliers. If all data points must be kept for

analysis, then more conservative algorithms should be used, e.g., LTSA or autoencoders [10,24,25].

2.7 Application to age estimation

2.7.1 Available data

We applied two NDLR techniques on a previously published dataset [16]. This dataset included clinical, dental and radiological data. Clinical data included geographical origin (country) and sex. Dental data were the eruption of the second and third molars (yes/no for each molar, i.e. 8 distinct variables). Radiological data included the readout of the forensic or radiologist expert regarding the fusion of the distal radius and ulna epiphyses. Additional data were: the alleged age of the person, the estimated age as provided by the Greulich and Pyle atlas [30], and the age estimated by the forensic examiner, on the basis of all the elements mentioned above. The alleged age was the age provided by the person examined. The radiological age was the age estimated based on the fusion of the distal radius and ulna epiphyses, according to the Greulich and Pyle atlas.

Countries were aggregated into 8 levels of one geographical origin variable (Africa, Asia, Western Europe, Eastern Europe, Middle East, Oceania, North America, South and Central America). Data were obtained over one year, from 1 January to 31 December 2007 in a suburban area near Paris, France. The age assessments were requested by the public prosecutor's office of Bobigny (Seine-Saint-Denis) for purposes of criminal or asylum proceedings. Examinations were conducted by trained forensic physicians.

2.7.2 Descriptive analyses

Median ages with 10^{th} and 90^{th} percentiles were computed for each case whether 2^{nd} or 3^{rd} molars were present or not; and whether gender was male or female. Differences in medians were assessed with Kruskal-Wallis test. Correlations between the age estimated by the forensic examiner, the skeletal age and the alleged age were also estimated.

2.7.3 Mapping techniques

We aimed at mapping every person for whom clinical and radiological data were complete in a low-dimensional space, so that the proximity or similarity of each pair of persons would be preserved. In such a mapping, if clinical and radiological data were representative of the age of a person, then two persons close to each other should have similar or close ages. Since data other than ages were categorical, we first applied a conservative transformation using multiple correspondence analysis (MCA), so that we obtained continuous variables. MCA is equivalent to PCA for discrete variables.

The second step consisted in estimating the intrinsic dimensionality of the dataset. We used the correlation dimension estimator. The third step consisted in applying two different NLDR techniques on these data, i.e. a conservative one, namely an autoencoder which is a kind of neural network [23-25], and a non-conservative one, the ISOMAP algorithm. The final step was the labeling of each person in the mapping provided by the NLDR techniques with their associated ages, i.e. respectively the alleged age, the estimated age according to the Greulich and Pyle atlas and the age estimated by the forensic physician, taken as a clinical, dental and radiological synthesis. So that figures could be readable, a random sample of 40 individuals out of the total number of available individuals was drawn and used for graphical display. MCA was performed with the statistical R software, and the ISOMAP and autoencoder algorithms run under MATLAB R2009b with the DR toolbox [31].

3. Results

3.1 A theoretical example

Results are reported in Figure 2. PCA was unable to unfold the Swiss roll dataset and behaved like a projection of data on a plane in a certain direction. ISOMAP behaved considerably better and was able to correctly unfold the dataset. In terms of computer time for execution, ISOMAP was time-consuming, since one of its steps consisted in applying a classical MDS algorithm, which was slow. Comparisons with other NLDR techniques and on other datasets (see supplementary material, Figures 1-3) all confirmed that NLDR methods perform better than PCA or MDS.

3.2 Age estimation data

3.2.1 Descriptive results

The initial dataset contained 499 people. Data were complete for 233 of them (46.7%). 74.7% were males, and the geographical origins were represented as follows: 96 from Africa (41%), 70 from Middle East (30%), 40 from Asia (17%), 18 from Western Europe (8%), 5 from South America, 3 from Western Europe and 1 from Oceania. The mean alleged age was 16.4 years (standard deviation SD: 2.1, median m [min-max]: 16.5 [9-36]). The mean estimated age according to the Greulich and Pyle atlas (resp. mean estimated age) was 17.8 (SD 1.58, m 18 [9-19.5]) (resp. 17.9 SD 2.08, m 18.5 [9-36]).

Table 3 presents the descriptive results. Individually, almost all variables were associated with significant differences in terms of median age as estimated by the forensic examiner. It was notably the case for the presence of 3rd molars. Estimated ages were strongly correlated to skeletal ages (Spearman's coefficient: 0.92, $p<0.001$) and alleged ages (0.73, $p<0.001$).

3.2.2 Results of the NLDR techniques for age estimation

The application of MCA on the data resulted in two principal axes accounting for 15.9% and 11.2%, respectively, of the total variance of the data. Visually, it appeared that the ages according to the Greulich and Pyle atlas, the alleged ages and the ages estimated by the forensic examiner were distributed along the principal axes at random.

According to the correlation dimension estimator, the intrinsic dimensionality of the dataset was 2. The results of the application of the ISOMAP and the autoencoder algorithms are displayed in Figures 3 and 4. The non conservative ISOMAP algorithm took into account 101 people out of 233, i.e. 132 were considered not connected to the 101 considered.

Starting from 13 variables routinely used to estimate age in clinical forensic settings, we ended up with an intrinsic data dimensionality of 2. A closer examination of the data obtained either with the ISOMAP or the autoencoder algorithms showed that in fact only 1 dimension seems to suffice to characterize data. Indeed, some areas of the 2 dimensional plane showed a linear arrangement of data points, suggesting that only one dimension may be sufficient to describe data .

In the 2-dimensional case, i.e. outside the particular subgroup of people that fitted a line, (figure 4), we found a high dispersion of the individuals across the plane, and no specific portion of this plane seemed to gather people according to a homogeneous age profile, e.g., a region of the plane gathering more specifically individuals aged between 14 and 17 years. Visually, the alleged ages, the radiological ages and the ages estimated as a synthesis of the forensic examination seemed to be randomly distributed across the plane.

3.2.3 Implications for the classification of individuals between adults and non-adults

The results yielded by both the Isomap and the autoencoder algorithms showed that ages under and above 18 years were intertwined with each other (data not shown). Otherwise stated, no straight or simple line seemed to be able to separate people younger than 18 from people older than 18 or give a clear linear and ordered trend for ages across the plane.

4. Discussion

We have presented the fundamental issues raised by the increase in dataset dimensionality, as well as the need to preserve the intrinsic data geometry as much as possible, in order to avoid mistakes in analysis. On one hand, it seems necessary to collect more and more data, but on the other, doing so will entail some unexpected pitfalls if we do not change our habits regarding conventional analysis. One way to cope with these issues might be to use NLDR techniques. Although NDLR techniques are recent, they have acquired some maturity and diversity, which allow making comparisons, especially with linear dimensionality reduction techniques. We can assume that they will be increasingly used in many fields, as has been shown with a few available examples [11,32,33] and forensic sciences should be no exception. Considering more data sources to estimate age will lead to bigger datasets that should be handled and analyzed carefully. Age estimation in living persons can be an opportunity to introduce more appropriate techniques like NLDR techniques.

These methods have limitations. First, like many other techniques, the user has to specify some parameters, which can be a matter of concern if there is no prior information on the intrinsic structure of data. However, only two parameters usually need to be specified, one of which is the estimated intrinsic dimension. Second, there is presently no consensus as to what the best intrinsic dimension estimator is for each situation and objective when these estimators can be used. The correlation dimension estimator is widely used and is usually recommended over other estimators, especially in physics, since it is correctly theoretically grounded [26,28]. The PCA eigenvalues estimator can only compute a dimension as an integer, i.e. 1, 2, 3… Unlike the other estimators we presented, it does not return a dimension between 1 and 2. For this reason, it is more prone to miss the correct dimension by one unit: if the correct dimension is 2, it could easily return 1, 2 or 3. Additionally, the value returned by this estimator is determined by the cut-off criterion for selecting the number of eigenvalues. Depending on this criterion, the returned value can also be rounded either up or down. We presented also the nearest neighbors dimension estimator. Although popular, this estimator is known for being flawed and unreliable: it is biased, sensitive to outliers and to edges and does not perform well, even in very simple examples [26]. Finally, there is still a lack of comparisons between NLDR techniques and classical techniques on real data. However, NLDR manifested its superiority on synthetic datasets [34,35] as well as in the first experiments on real data [32-34].

The absence of documented chronological ages is a major limitation of our forensic study and prevents our method to be fully validated on these data. Since many NLDR techniques enable the construction of maps that locate individuals according to their similarities, a first step to further standardize and validate our approach would be to check that individuals sharing both the same characteristics and the same documented chronological age are closer to each other than to individuals with different characteristics. This should be executed in the most controlled way possible, which means using data uniformly and evenly distributed in terms of each characteristic (the same number of individuals for each age, for each geographical origin, for each gender, and so on), all examinations and measurements being performed according to the same protocol. For the classifying problem (adult vs non-adult), the sample to consider should be narrower in age ranges (for example: 10 to 25 years), with more precise age data (expressed in months rather than in years, especially around 18 years of age). Approximatively about 500 individuals of each gender and origin should be considered if age is given in months (10 individuals of each age) and about 125 of each gender and origin if age is given in trimester. For both cases (age estimation and adulthood estimation), it seems important to consider the alleged age since this available information cannot be completely ignored and the patient's voice unheard. Obviously, the main obstacle remains the difficulty – not to say the impossibility - to rely on documented chronological age for migrants who are usually subjects to age estimation procedures.

In our forensic age estimation study, clinical, dental and radiological data that are routinely used to estimate one person's age failed to sort out people so that their similarity according to these characteristics translates into similar estimated ages. It appeared that a problem described by 13 distinct variables was collapsible to a one- or two-dimensional problem,

although the classical approach, i.e. with MCA, suggested that two dimensions only accounted for 27.1% of the total variance. This suggests a high redundancy – or equivalently a poor informational value – of the data. Moreover, the locally one-dimensional representation of data implies the possibility of a perfect linear regression. Unfortunately, figure 3 demonstrated that if this part of data could be depicted by a line, it failed to sort out the different ages in a correct order. This combination of clinical, dental and radiological data could not explain the ages that we registered. Worse, their addition or integration seemed to be useless. It therefore questions the relevance of searching for more variables to integrate and compare. This new insight to the age estimation problem highlights the fact that the existence of a linear correlation between some characteristics such as radiological features and the chronological age is one kind of evidence, but it has limited value for the accurate estimation of one particular person's age. Besides, explaining and predicting facts are distinct, rarely compatible tasks [36]. Our findings do not contradict that such a correlation exists. They merely illustrate that this correlation is not an efficient way to estimate the age of a person or to decide whether a person is less or more than 18 years old.

The goals we aim at – estimating the age, classifying persons with respect to being above 18 years – must not be confounded with our will to understand as accurately as possible the physiology of ageing. Even high standards in molecular analysis have failed short to predict a physiological age more accurately than a characteristic time of one year [37]. If integrating data of different natures to estimate age is laudable, the linear regression techniques used so far [19] are inappropriate since they are above all explanatory techniques (i.e. they distribute the overall variance of data among the variables of interest according to their respective contribution to this variance). They are in no way predictive methods although they are abusively called so. They are useful in a risk factor approach, to identify and quantify independent factor risks. Such modeling approaches cannot be satisfactory in age estimation unless they are very accurate. In the case of deciding whether a person is above 18 years old, we should focus on the best available techniques that present the best performances in terms of classification, and try them on available data. Highly effective techniques exist that should be challenged in the field of forensics, such as the Support Vector Machine (SVM) algorithms [24]. However, such techniques require the previous knowledge of the real chronological age of a subsample of the people of whom we want to estimate the age. The performances of SVM algorithms far exceed the results that can be expected from linear techniques, such as regression techniques previously presented in age estimation [9] and have been acknowledged in various settings [38,39]. So far, they are gaining a wider acceptance in clinical fields, particularly for improving diagnostic tests [39].

In this suggested way to improve performances in age estimation, not knowing the real age of the examined individuals is a limitation to our work. Therefore, we strongly encourage researchers to duplicate our findings on their own data. Similarly, we had no basic characteristics such as weight and height, nor more sophisticated imaging methods, such as clavicle CT scans or Magnetic Resonance Imaging [40,41]. There is nonetheless another way to cope with age estimation in the case where the real age remains unknown, which can be provided by Bayesian approaches [42,43]. Recently, they also proved promising for classifying individuals based on dental evidence and relying on soft evidence [44,45]. Moreover, Bayesian approaches could take into account the age alleged by the person, which would provide a more ethical approach to this problem. However, whether SVM or Bayesian approaches are chosen, both cases require that researchers gain confidence into these now well-known techniques, can handle and criticize them, and if they prove efficient, use them in their daily practice.

5. Conclusion

The integration of various sources of information to improve accuracy in estimating the age of living persons may be considered cautiously and in accordance to the goal we aim to: estimating the age or classifying persons according to a threshold age. The increase of data

amounts present specific issues that a forensic scientist should be aware of and that must be dealt with using adequate techniques.

Competing interests: We declare that we have no conflicts of interest

Funding: None

List of Abbreviations:
HLLE: Hessian LLE
ISOMAP: Isometric mapping
LLE: Locally Linear Embedding
LTSA: Local Tangent Space Alignment
MCA: Multiple correspondences analysis
MDS: Multidimensional scaling
NLDR: Nonlinear dimensionality reduction
PCA: Principal components analysis.
SVM: Support Vector Machine

References

1 Galea S, Riddle M, Kaplan GA. (2009) Causal thinking and complex system approaches in epidemiology. Int J Epidemiol 39:97–106

2 Reshef DN, Reshef YA, Finucane HK, et al. (2011) Detecting Novel Associations in Large Data Sets. Science 334:1518–1524

3 Connolly AC, Guntupalli JS, Gors J, et al. (2012) The representation of biological classes in the human brain. J Neurosci 32:2608–2618

4 Priyadarshini G, Sarmah R, Chakraborty B, et al. (2012) An effective graph-based clustering technique to identify coherent patterns from gene expression data. Int J Bioinform Res Appl 8:18–37

5 Conry MC, Morgan K, Curry P, et al. (2011) The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC Public Health 11:692

6 ENCODE Project Consortium, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

7 Bellman R. (2003) Dynamic programming. Dover Publications, Mineola

8 Lee JA, Verleysen M. (2007) Nonlinear dimensionality reduction. Springer, Berlin

9 Tenenbaum JB, de Silva V, Langford JC. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290:2319–2323

10 Izenman AJ. (2008) Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, illustrated edition. Springer-Verlag, New York

11 Gorban AN, Zinovyev A. (2010) Principal manifolds and graphs in practice: from molecular biology to dynamical systems. International Journal of Neural Systems 20:219–232

12 Olze A, Reisinger W, Geserick G, Schmeling A. (2006) Age estimation of unaccompanied minors. Part II. Dental aspects. Forensic Sci Int 159 Suppl 1:S65–67

13 Baumann U, Schulz R, Reisinger W, Heinecke A, Schmeling A, Schmidt S. (2009) Reference study on the time frame for ossification of the distal radius and ulnar epiphyses on the hand radiograph. Forensic Sci Int 191:15–18

14 Mansourvar M, Ismail MA, Raj RG, et al. (2014) The applicability of Greulich and Pyle atlas to assess skeletal age for four ethnic groups. J Forensic Leg Med 22:26–29

15 Bassed RB. (2012) Advances in forensic age estimation. Forensic Science, Medicine, and Pathology 8:194–196

16 Pruvost MO, Boraud C, Chariot P. (2010) Skeletal age determination in adolescents involved in judicial procedures: from evidence-based principles to medical practice. J Med Ethics 36:71–74

17 Schulz R, Schiborr M, Pfeiffer H, Schmidt S, Schmeling A. (2014) Forensic age estimation in living subjects based on ultrasound examination of the ossification of the olecranon. J Forensic Leg Med 22:68–72

18 Manzoor Mughal A, Hassan N, Ahmed A. (2014) Bone Age Assessment Methods: A Critical Review. Pak J Med Sci 30:211–215

19 Bassed RB, Briggs C, Drummer OH. (2011) Age estimation using CT imaging of the third molar tooth, the medial clavicular epiphysis, and the spheno-occipital synchondrosis: a multifactorial approach. Forensic Sci Int 212:273.e1–5

20 Lovejoy CO, Meindl RS, Mensforth RP, and Barton TJ. (1985) Multifactorial determination of skeletal age at death: A method and blind tests of its accuracy. Am J Phys Anthropol 68:1-14

21 Robinson A. (1988) American Cartographic Association. Committee on Map Projections. Choosing a world map: attributes, distortions, classes, aspects. Falls Church

22 Pearson K. (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559–572

23 Hinton GE, Salakhutdinov RR. (2006) Reducing the Dimensionality of Data with Neural Networks. Science 313:504–507

24 Bengio Y, Courville A, Vincent P. (2013) Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828

25 Bengio Y, Yao L, Alain G, Vincent P. (2013) Generalized Denoising Auto-Encoders as Generative Models. arXiv:13056663 [cs] 2013. (accessed 12 Jan 2015)

26 Camastra F. (2003) Data Dimensionality Estimation Methods: A Survey. Pattern Recognition 36:2945–2954

27 Carter KM, Member S, Raich R, Iii AOH. (2010) On Local Intrinsic Dimension Estimation and Its Applications. IEEE Trans Signal Process 58:650–663

28 Grassberger P, Procaccia I. (1983) Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena 9:189–208

29 Levina E, Bickel PJ. (2005) Maximum Likelihood Estimation of Intrinsic Dimension. Advances in Neural Information Processing Systems 17:777–784

30 Greulich WW, Pyle SI. (1959) Radiographic atlas of skeletal development of the hand and wrist, Stanford: Stanford University

31 Matlab Toolbox for Dimensionality Reduction, v0.8.1, 2013, http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html. (accessed 12 Jan 2015)

32 Shi J, Luo Z. (2010) Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Computers in Biology and Medicine 40:723–732

33 Weng S, Zhang C, Lin Z, Zhang X. (2005) Mining the structural knowledge of high-dimensional medical data using isomap. Med Biol Eng Comput 43:410–412

34 Gorban AN. (2008) Principal manifolds for data visualization and dimension reduction. UK Springer London, London

35 Van der Maaten LJP, Postma EO, van den Herik HJ. (2009) Dimensionality Reduction: A Comparative Review. Tilburg, Tilburg University Technical Report

36 Thom R. (1999) Prédire n'est pas expliquer. Flammarion, Paris

37 Gibbs WW. (2014) Biomarkers and ageing: The clock-watcher. Nature 508:168–170

38 Howe A, Escalona OJ, Di Maio R, et al. (2014) A support vector machine for predicting defibrillation outcomes from waveform metrics. Resuscitation 85:343-349

39 Garcia Molina JF, Zheng L, Sertdemir, et al. (2014) Incremental learning with SVM for multimodal classification of prostatic adenocarcinoma. PLoS One 9:e93600

40 Krämer JA, Schmidt S, Jürgens Ku, et al. (2014) Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med 128:509-514

41 Wittschieber D, Schulz R, Vieth V, et al. (2014) Influence of the examiner's qualification and sources of error during stage determination of the medial clavicular epiphysis by means of computed tomography. Int J Legal Med 128:183–191

42 Pe'er D. (2005) Bayesian network analysis of signaling networks: a primer. Sci STKE, 26;2005(281):pl4

43 Chariot P, Caussinus H. (2015) Age estimation in undocumented migrant adolescents: medical response to judicial authorities. Presse Med 44:99-100

44 Corradi F, Pinchi V, Barsanti I, et al. (2013) Optimal age classification of young individuals based on dental evidence in civil and criminal proceedings. Int J Legal Med 127:1157-1164

45 Corradi F, Pinchi V, Barsanti I, Garatti S. (2013) Probabilistic classification of age by third molar development: the use of soft evidence. J Forensic Sci. 58:51-59

Table 1: The three steps of the ISOMAP algorithm

Table 2: Intrinsic dimensionality estimators applied to four theoretical examples.

GMST: geodesic minimum spanning tree. Correlation dimension and maximum likelihood estimators succeed in approaching the intrinsic dimension in most, if not all, the examples. Estimators are described in [25,27,28,S2-S4]

A dimension is not necessarily an integer, it can be fractal [S5,S6]. Lines drawn on a sheet of paper can present qualitative differences in terms of their aspects, and yet they share the common property of being one-dimensional objects. Whether a specific point has to be

located on a straight line or on a line consisting of many zigzags, only one unique coordinate is necessary because these two lines are both one-dimensional. The straight line will not fill space significantly, while the zigzag line will occupy more space. It is then possible to define a fractal dimension to characterize these two lines, which will be between 1 and 2, depending on their ability to fill 2-dimensional space.

Table 3: Median ages estimated by the forensic examiner, for each variable (gender, 2nd and 3rd molars).

*: p<0.01, **: p<0.001. Median age estimated by the forensic examiner is given for each case whether a 2nd or 3rd molar is present or absent, with the 10th and 90th percentiles into brackets and whether the person is male or female.

Legends of figures

Figure 1: Distances over two different spaces: flat space and curved space.

Changes in proximity relationships for a, b and c. Distances that appear to be correct and relevant in the case of a nonstructured space (case A) will turn out to be incorrect if the space is structured and is described by a lower intrinsic dimension (case B): cars cannot go through buildings; they have to use roads.

In case A, the closest person to “b” is “c” because the ambient space is flat (i.e., 2-dimensional), with no specific structure. In case B, the closest person to “b” is “a” because data are intrinsically structured as a spiral. Considering that “b” and “c” belong to the same group on the basis of their proximity leads to a wrong statement if it is measured the same way in both cases. Proximity in case B should be measured in the way in which the dot style distance approximates it. By definition, a distance that measures the shortest path between two points belonging to a specific curved space is called a geodesic distance. In case A, we need a 2-dimensional space to describe data. Two coordinates are needed to locate someone. In case B, when we respect the intrinsic spiral geometry, we only need one degree of freedom: it is merely a line with curvature, and only one coordinate along that line is enough to locate someone.

Figure 2: A theoretical example (Swiss roll) and its expected unfolding, in which a NLDR technique (ISOMAP) is compared with PCA and MDS.

Top: Example of the Swiss roll (left) and its expected unfolding (right)

Bottom: PCA (left), MDS (middle) and ISOMAP (right) unfoldings. Execution times are given in seconds or minutes.

Figure 3: The ISOMAP algorithm applied to age estimation: radiological ages and ages estimated by the forensic examiner are seemingly randomly distributed along a line.

The ISOMAP algorithm identified 101 individuals significantly connected with each other out of 233. The more similar two individuals are in terms of clinical, dental and radiological data, the closer they are to each other. Despite this, two identical or similar individuals can have rather different radiological ages (top) or different ages as estimated by the forensic examiner (bottom). No order seems to be identified along the line. 40 individuals out of 101 have been randomly chosen and reported here.

Figure 4: The autoencoder algorithm applied to age estimation: alleged ages (A), radiological ages (B) and ages estimated by the forensic examiner (C) are seemingly randomly distributed across the whole plane.

The autoencoder algorithm is a conservative algorithm that keeps all 233 initial individuals and dispatches them across a 2-dimensional space that preserves their topological relationships. The more similar two individuals are in terms of clinical, dental and radiological data, the closer they are to each other. Despite this, two identical or similar individuals can have rather different alleged ages (A). No specific age clustering seems to prevail across the whole plane. The same observation can be done for radiological (B) or forensic estimated ages (C). 50 individuals out of 233 have been randomly chosen and reported here. Only selected points are labelled to ensure readability.

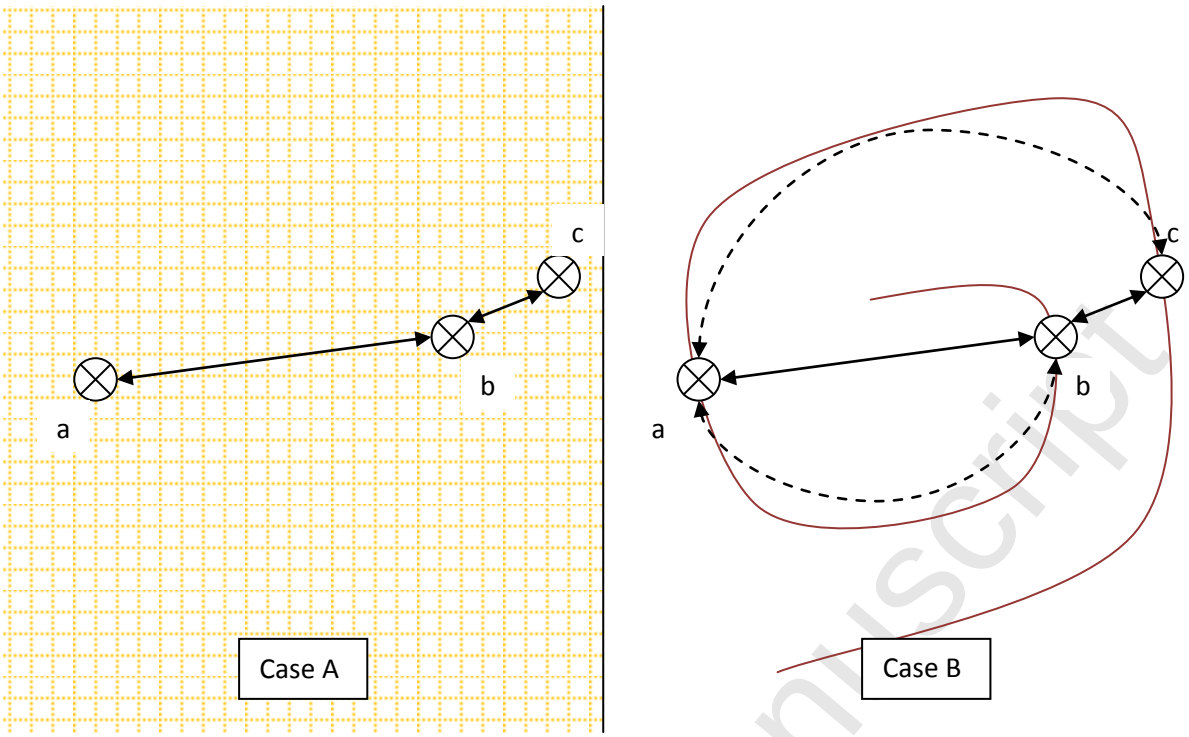
Steps	Description
Step 1: construction of a neighborhood graph	For each data point, the algorithm builds the graph (i.e. the connection) between it and its *k*-nearest neighbors. The graph is distance-weighted. This means that each edge is associated with a Euclidean distance, which is measured between the two points (also called “vertices”).
Step 2: computation of geodesic paths for the purpose of building the geodesic distances matrix:	For each pair of points, the geodesic distance is approximated in two ways. For direct neighbors, the geodesic distance is approximated by the Euclidean distance. Otherwise, the shortest path between two points is computed through a classical graph algorithm, such as the Dijkstra algorithm.
Step 3: use of a classical multidimensional scaling (MDS) algorithm to reduce dimensionality:[S1]	The MDS is run on the geodesic distance matrix and delivers the dimensionally reduced dataset.

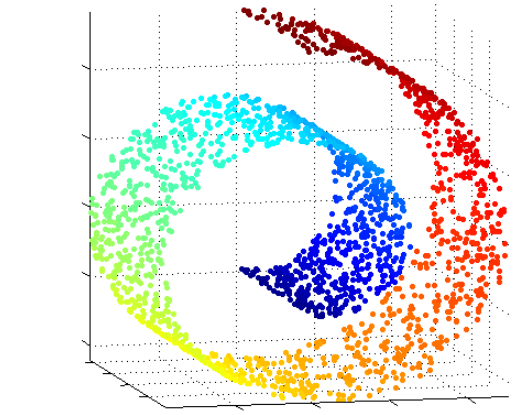
	Dataset examples			
	Swiss roll	Toroïdal helix	Twin peaks	Infinite
Intrinsic dimension to retrieve	2	1	2	2
Estimators :				
Correlation dimension	1.94	1.47	2.02	2.29
Nearest neighbor dimension	0.56	0.7	0.51	0.44
GMST dimension	1.77	1.38	2.57	2.5
Packing numbers dimension	2.22	1.11	1.31	0.91
PCA eigenvalues dimension	2	2	2	2
Maximum likelihood dimension	1.94	1.5	2.14	2.53

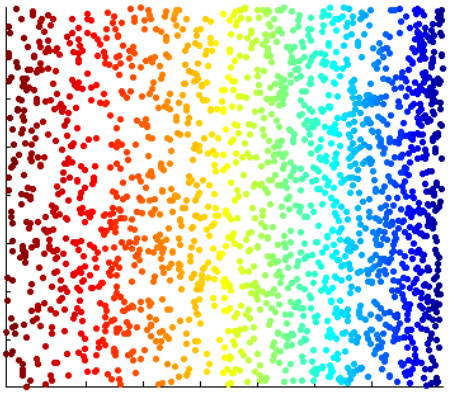
2nd molars	17**	27**	37*	47*
Present	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.0 [15.5; 19.0]
Absent	15.8 [10.5; 19.0]	15.5 [10.0; 17.5]	17.0 [11.0; 19.0]	16.0 [10.0; 19.0]

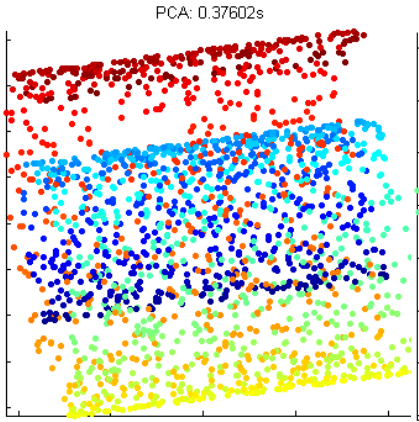
3rd molars	18**	28**	38**	48**
Present	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]
Absent	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.3 [14.5; 19.0]

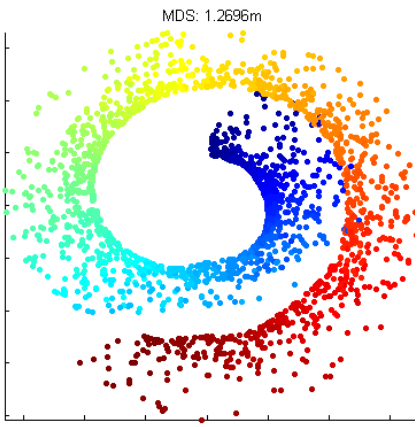
Gender**	
Male	19.0 [15.5; 19.0]
Female	17.5 [15.0; 19.0]

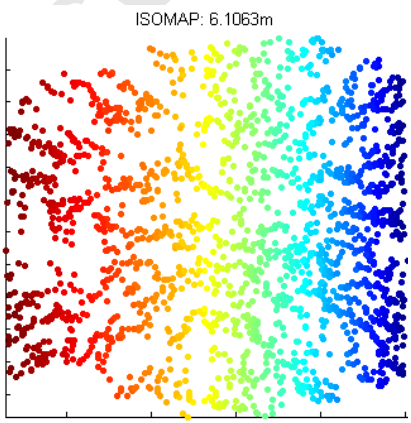


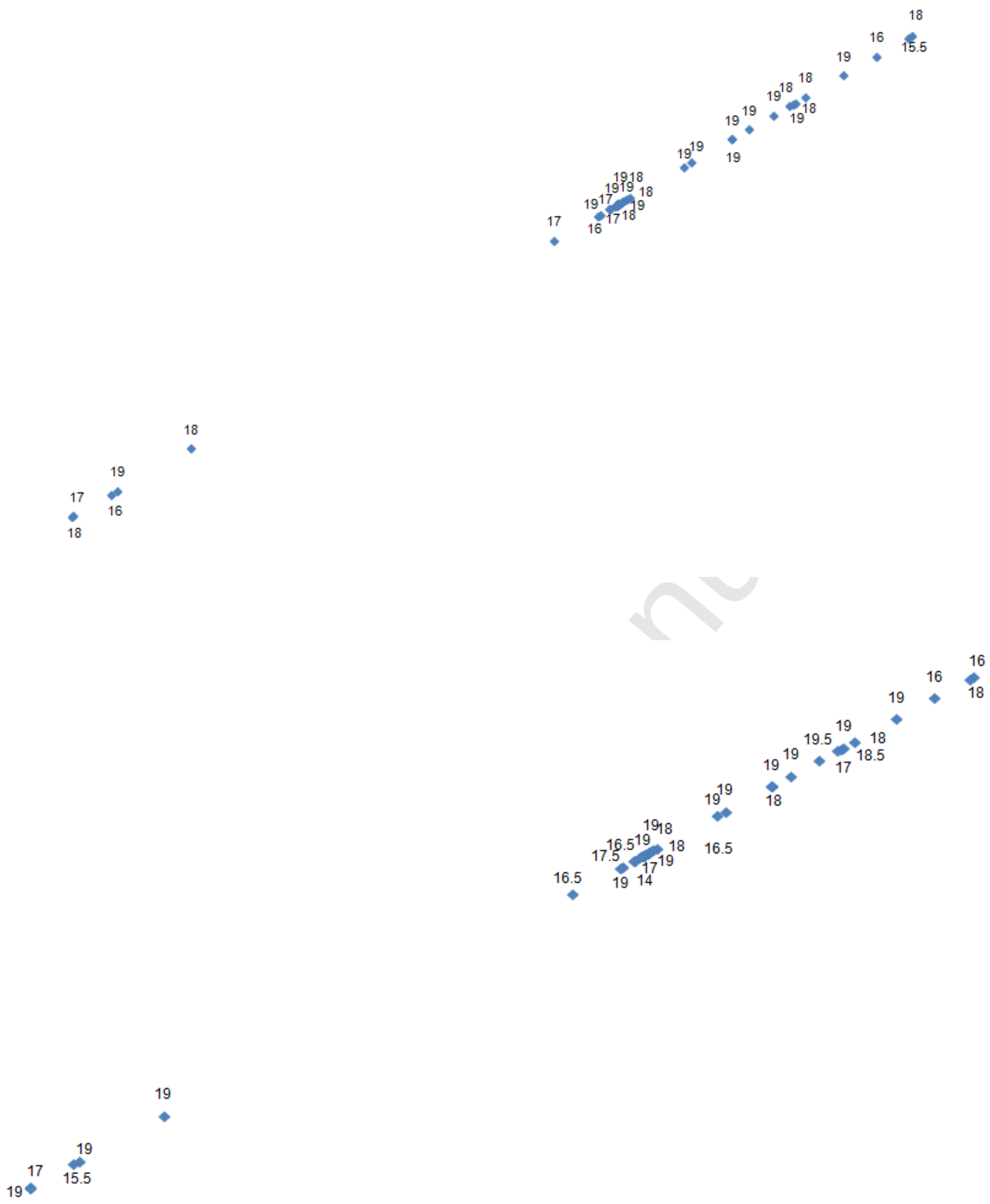


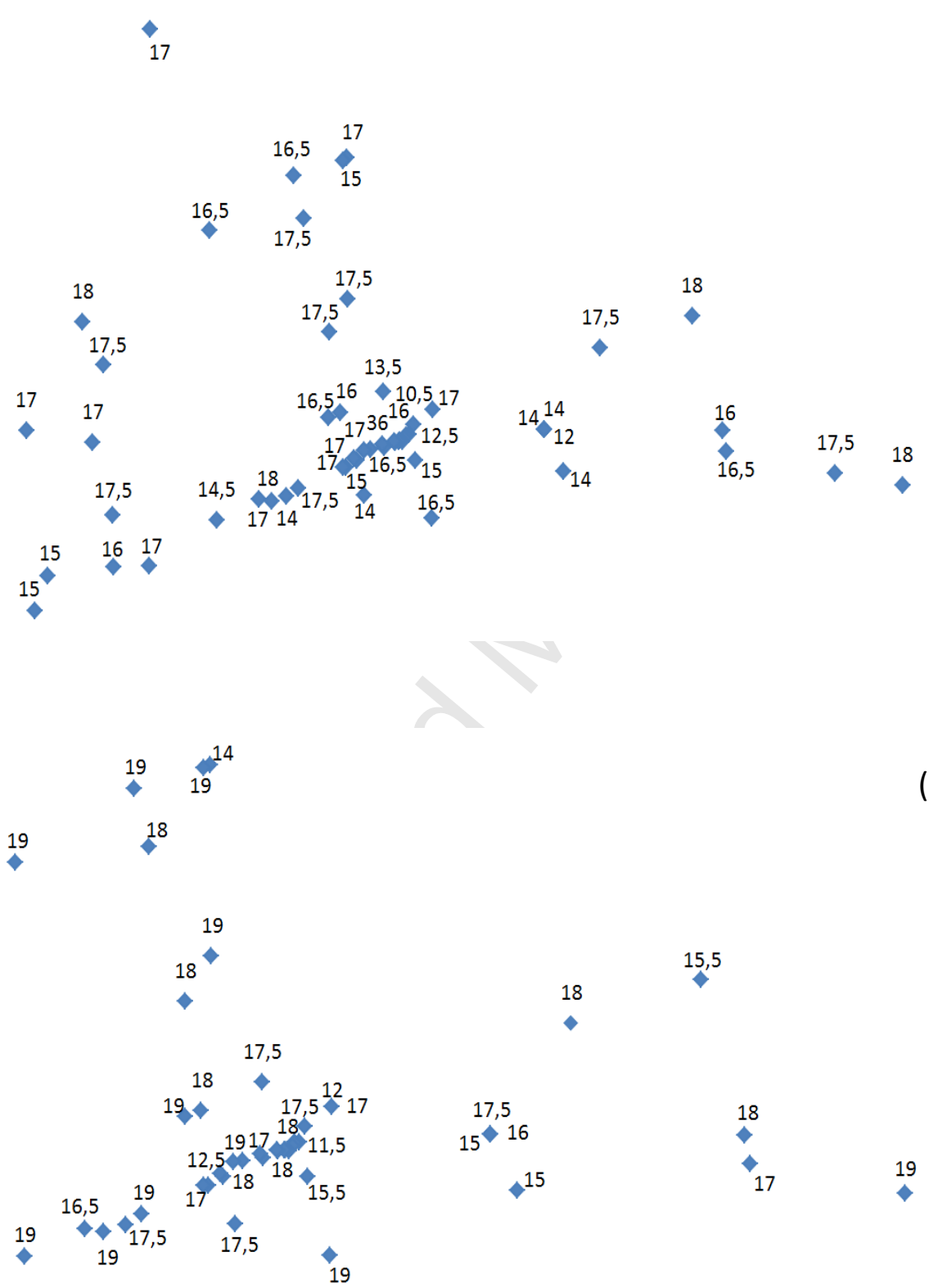












(C)