Linear-Time Hierarchical Matching and Regression: Application to SDSS Spectroscopic and Photometric Redshifts

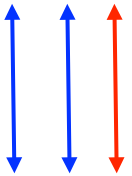
p-Adic regression
Ultrametric wavelet regression

Fionn Murtagh, IAU 325 – Astro2016

Baire, or longest common prefix distance. Also an ultrametric.

An example of Baire distance for two numbers (x and y) using a precision of 3:

$x = 0.425$



$y = 0.427$

Baire distance between x and y:

$d_{\mathcal{B}}(x, y) = 10^{-2}$

Base ($\mathcal{B}$) here is 10 (suitable for real values)

Precision here = |K| = 3

That is:

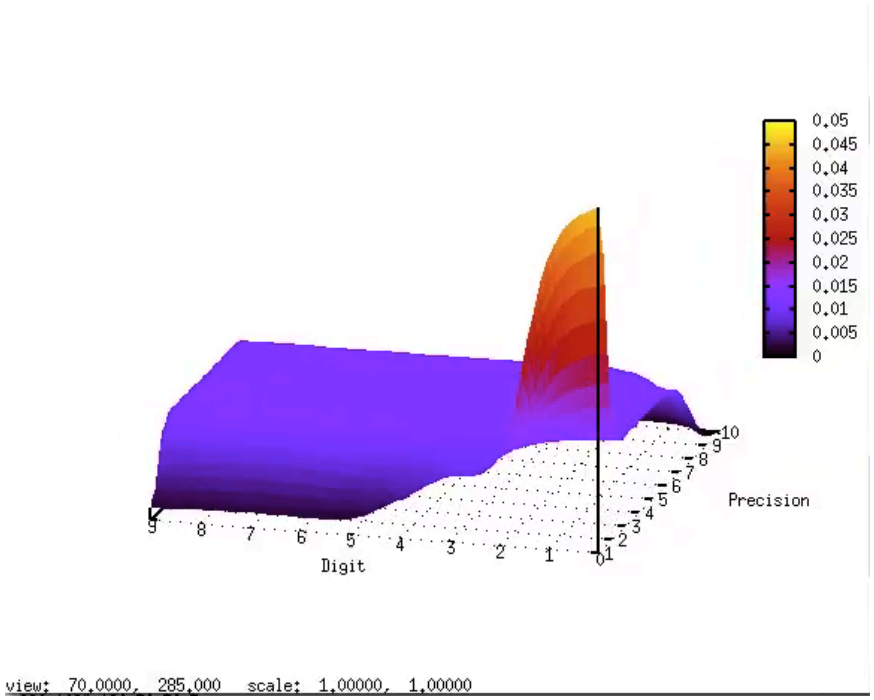
k=1 -> $x_k = y_k$ -> 4

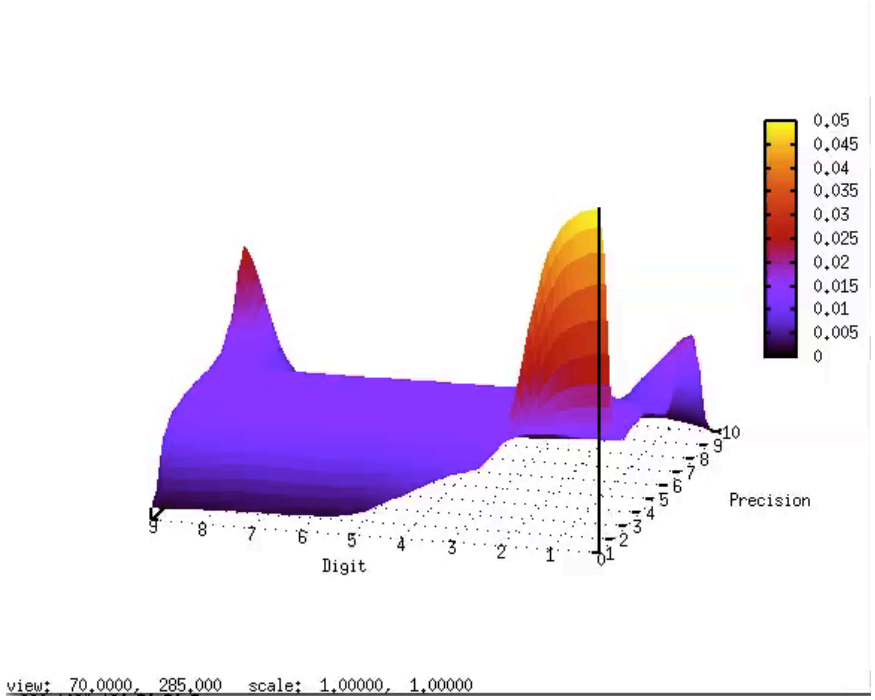
k=2 -> $x_k = y_k$ -> 2

k=3 -> $x_k \neq y_k$ -> 5≠7

- Approx. 0.5 million SDSS release 5 (D'Abrusco et al.) - regress z_spect on z_phot
- Furthermore: determine good quality mappings of z_spect onto z_phot, and less qood quality mappings
- I.e., cluster-wise nearest neighbour regression
- Note: cluster-wise not spatially (RA, Dec) but rather within the data itself

Perspective Plots of Digit Distributions





- On the right we have z_spec where three data peaks can be observed. On the left we have z_phot where only one data peak can be seen.

Outcome of Fast Clusterwise Regression - using
m-adic hierarchical clustering, i.e. Baire distance with base 10.

- 82.8% of z_spec and z_phot have at least 2 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3, 2 decimal digits.
- We can find very efficiently where these 82.8% of the astronomical objects are.
- 21.7% of z_spec and z_phot have at least 3 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3 decimal digits.

Note on terminology, m-adic (or m-ary), p-adic number systems

- Notationally m >= 2 is a positive integer, and p is prime.
- (m-adic when m = 10 is decimal. p-Adic when p = 2 is binary, when p = 3 is ternary.)
- An m-adic number system is a ring, while a p-adic number system constitutes a field. A field has a multiplicative inverse for non-zero values, i.e. it permits division.
- By convention, p-adic for p = ∞ is real. For reals, e.g. 0.5 is identical to 0.49999….
- In an m-adic representation, for m = 10, we distinguish 0.50 from 0.49, and even 0.5 from 0.49999...
- Consider the tree representation of an m-adic number. E.g. for m = 10, this is a regular 10-way tree.
- Rather than real number valued proximity, i.e. similarity of tree terminal nodes, our focus will be on the matching of branches, from the root, of the tree representation.

Introductory motivation

- Use heterogeneity and diversity in data. An a priori and global model may not be appropriate.
- Concerned with matching, and drawing inferences (extrapolation and interpolation, prediction, distributional degree of association, etc.) from structures that are discrete.
- In addition to being discrete, there are associations, similarities and identities that are relevant.
- Also relevant are incorporation, inclusion, properties of an object being a subset of properties of one or more other objects.

- A representation for such structures is an ultrametric or tree topology.
- Objects are taken as nodes of a tree. A tree is a synonym for hierarchy.
- These objects, or entities, could, if desired, include sub-objects and sub-entities also.
- In set notation, a hierarchy is a partially ordered set, or poset

Short review of other work, 1/3: Determining Photometric Redshifts from Colour and Magnitude Observed Data, and Evaluating relative to Spectroscopic Redshifts

- In Vanzella et al., “Photometric redshifts with the multilayer perceptron neural network: application to the HDF-S and SDSS”, A&A 423, 761-776, 2004, there is predicting of photometric redshifts “from an ultra deep multicolor catalog''. Training is carried out with spectroscopic redshifts. This is noted: “the difficulty in obtaining spectroscopic redshifts of faint objects”, and then: “A crucial test in all cases is the comparison between the photometric and spectroscopic redshifts which is typically limited to a subsample of relatively bright objects''.
- Csabai et al., “Photometric redshifts for the SDSS Early Data Release”, AJ 125, 580-592, 2003.
- Firth et al., “Estimating photometric redshifts with artificial neural networks”, MNRAS 339, 1195-1202, 2003.

Short review 2/3: Interval Measurements for Bayesian "stacking" Modelling; Accuracy and Correctness of Measurement

- In Shu et al., "Evolution of the velocity-dispersion functino of luminous red galaxies: a hierarchical Bayesian measurement", AJ 124, 90-100, 2012, velocity distributions are at issue, for association with galaxy sizes, to "determine 'dynamical masses' that are independent of stellar-population assumptions", with that to be used for evolution of galaxies for given mass, following relationship estimation with mass and gravitational potential. Interest is in elliptical galaxies, that are "To a first approximation ... 'pressure-supported' rather than rotationally supported". Velocity dispersion is to be based on spectroscopic data.

- Now, in particular for faint, even if luminous, galaxies, there will be uncertainty and non-Gaussianity in measurement. Eigenspectra are determined from principal components analysis. Because of the imprecision of measurement the following is carried out, in the estimation of velocity dispersion.

- Both in redshift and in absolute magnitude, respectively with intervals of 0.04 and 0.1, galaxies are binned. Therefore, for error or imprecision of measurement, binning, i.e. interval measurements, are a way to somewhat robustify the data. Based on extensive analyses, it is concluded that here the "stacking" of multiple spectra is replaced by a new "Bayesian stacking" approach. (A hierarchical Bayesian approach.)

- Bolton et al., "Spectral classification and redshift measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey", AJ 144, 144-164, 2012.
- Under "Known issues", there are the following: the use of probability priors on principal component analysis coefficient combinations; spectra that are obscured by others, e.g. quasar spectra, by AGN spectra; spectra affected by "cross-talk from bright stars"; superpositioning of observed objects; and a few class of object, and detector suitability ("fibers near the edge of the spectrograph camera fields of view")

Short review 3/3: Nonlinear regression

- R. d'Abrusco, G. Longo et al, "The use of neural networks to probe the structure of the nearby universe", Prof. ADA-4, 2006, arXiv Jan. 2007: MLPs are used to relate photometric redshifts to spectral information. Varying object classes (normal galaxies, stars, late type stars, nearby AGNs, distant AGNs) are subject to PCA of spectra, to provide an eigenvector-based spectral classification index.
- The case is then made for carrying out the nonlinear regression, using MLP, on two different redshift intervals, $z < 0.25$ and $z > 0.25$. Differing galaxy populations are associated with these redshift intervals.
- R. d'Abrusco et al., "Mining the SDSS archive. I. Photometric redshifts in the nearby universe", AJ 663, 752-764, 2007.
- "photometric redshift estimates depend on the morphological type, age, metallicity, dust, etc. it has to be expected that if some morphological parameters are taken into account besides than magnitudes or colors alone, estimates of photometric redshifts should become more accurate." In this work, the "near universe'', $z < 0.5$, is at issue, and also with discussion of "the near and intermediate redshift universe", $z < 1$.
- "the derivation of photometric redshifts requires, besides an accurate evaluation of the errors, also the identification of a homogeneous sample of objects."

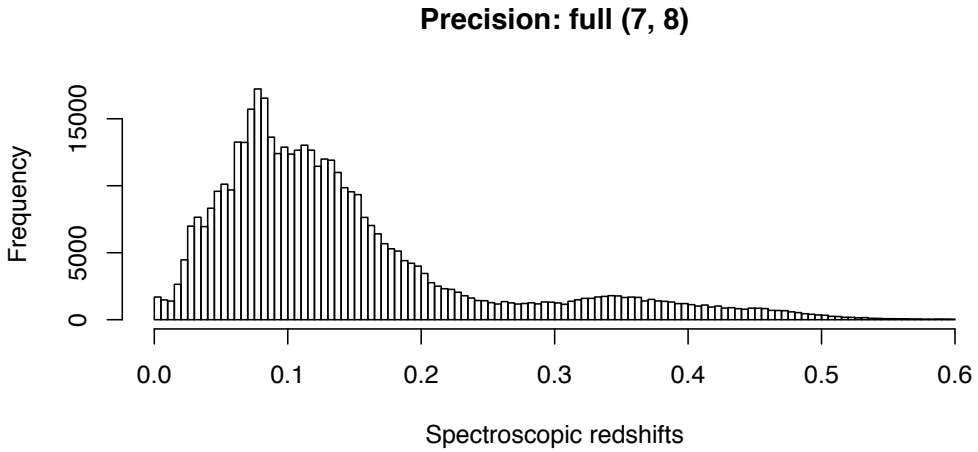
Data and Objectives Pursued Here

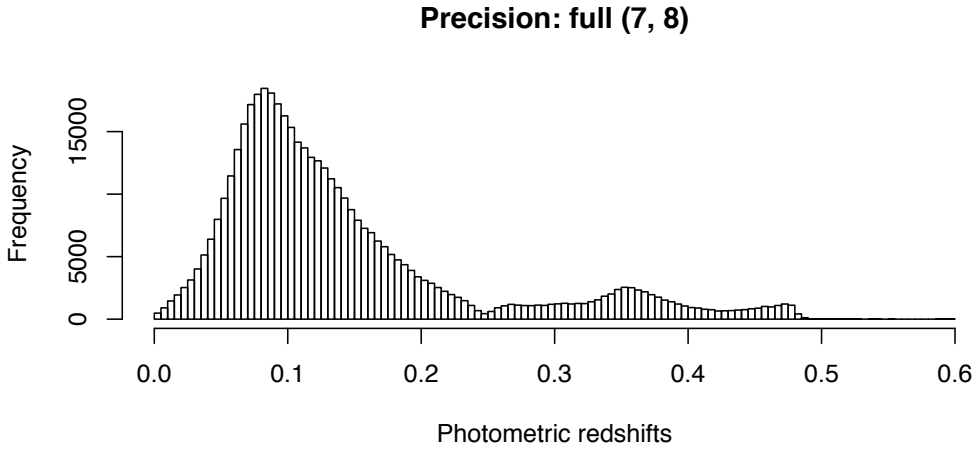
- SDSS Data Release 5, relating to the following: “Stripe 82 is an equatorial region repeatedly imaged during 2005, 2006, and 2007’’.
- Number of objects: 443094; right ascension, declination, spectroscopic redshift, photometric redshift. Then minimum redshifts, respectively spectroscopic and photometric, are: 0.000100049, 0.0001035912, and the maximum redshifts are: 0.599886, 0.5961629.

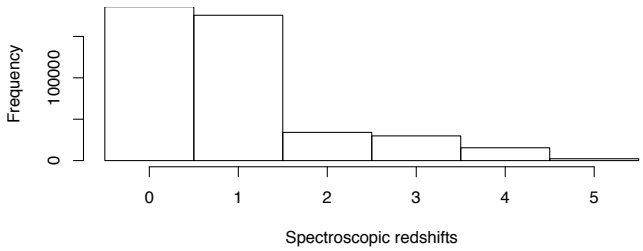
- Assess spectroscopic redshift from photometric redshift:
- Take discreteness of measurement into account.
- Take distinction of value to be primarily associated with the discrete sourcing of our measurements, rather than being solely a statistical uncertainty or error component of our measurement.
- However statistical uncertainty or error component of measurement are taken as integral to the discreteness of sourced data.
- It arises from this reasoning that what is important in practice is to be able to codify one's data, in the sense both of data encoding and of data representation, here related to number theory.
- From the data encoding and representation, we are seeking to associate data interpretation and understanding, with the discrete sourcing of our measured data.

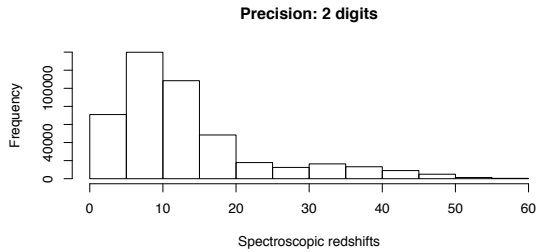
- *Preliminary exploratory phase of analysis follows.*

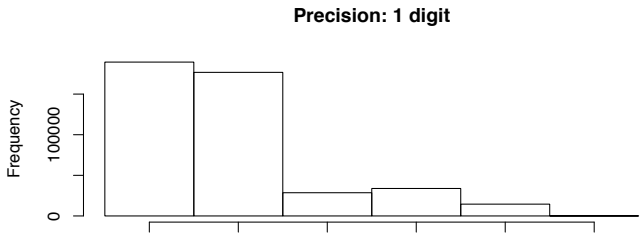
Full precisions, 7 or 8 digits. While mainly peaked around lower redshifts, there are some potentially interesting smaller peaks.

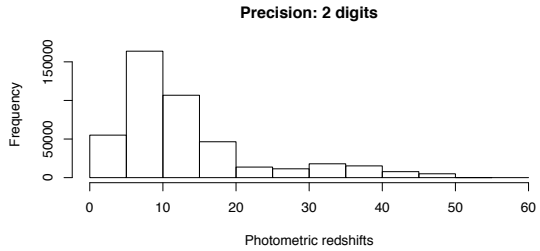


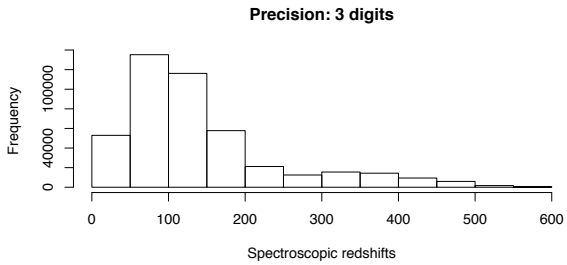


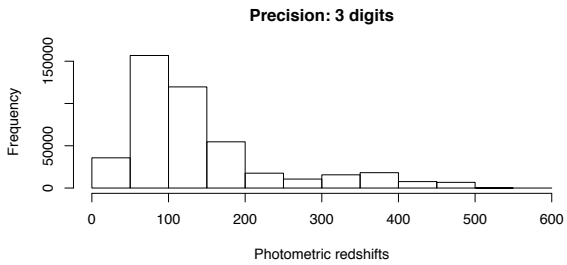












Histograms of (upper) spectroscopic, (lower) photometric redshift values.
Upper left: 1 digit precision.
Upper right: 2 digits precision
Lower left: 3 digits precision

We are considering the histograms as a preliminary exploratory phase of the analysis

- Overall, the first digit of precision of the spectroscopic and photometric redshifts is common to 366907 objects, that is, 83% of all cases. Can this be furthered?

- If we look at both the shared first digit of precision, and additionally a difference in the first digit of precision of at most 1, then we find that 99.6% of all the spectroscopic and photometric redshift measurement are that close in measurement value.
- While this is motivational, it requires further study of just what redshifts differ by 1 in the first digit of precision. However, we do not consider such a finding as generally and broadly applicable.

Re-Representing Our Data in p-Adic and Other Number Systems

- With reference to the histograms displayed in 3-dimensional Euclidean space, in regression-oriented matching, we could “calibrate” the regression with RA and Dec.
- F. Murtagh, A.E. Raftery and J.L. Starck, “Bayesian inference for multiband image segmentation via model-based cluster trees", Image and Vision Computing, 23, 587-596, 2005.

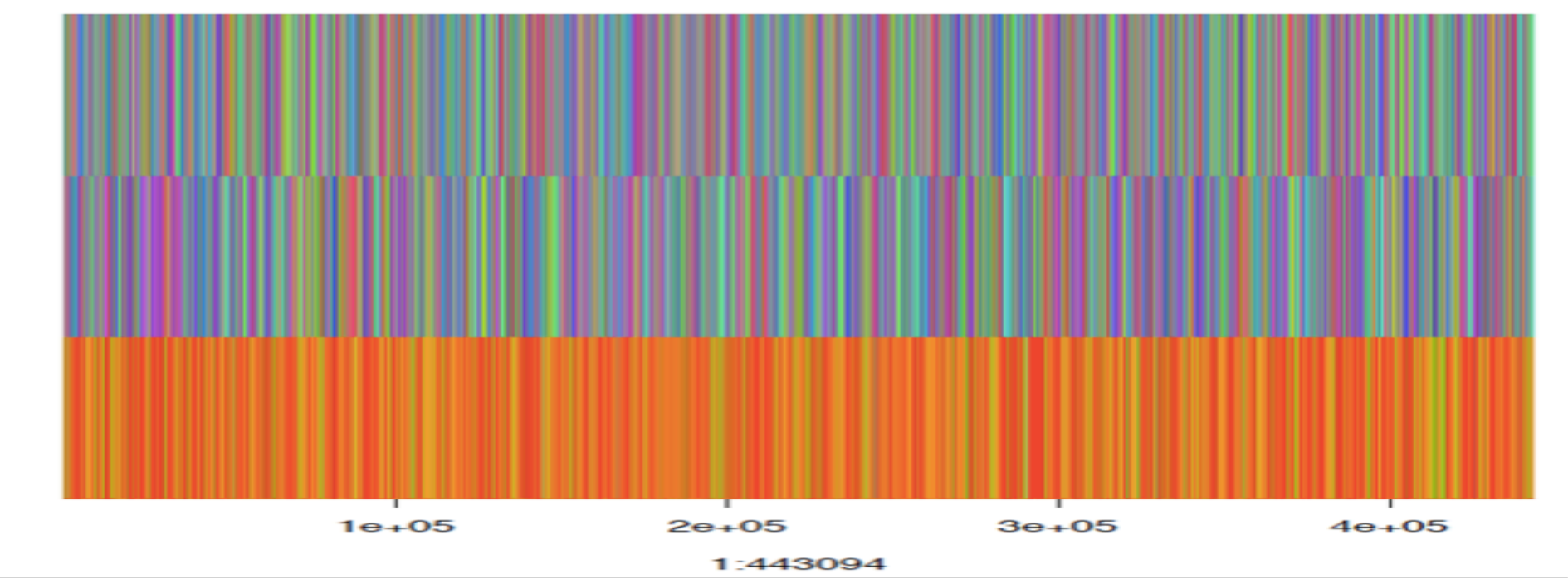
- Now, compared to a Euclidean and Hilbert space, we are dealing with discrete object locations, and clustered albeit delimited regions of objects. A graph and more particularly, a tree is an appropriate representation, rather than a continuous space.
- Because of the directly mapped, rooted tree representation that can be associated with any m-adic number representation, we proceed as follows: consider our given decimal or base 10 measurements, as m-adic with m = 10. Efficiently derive other m-adic number representations, to assess them.

Re-Representing Data in Other Number Systems, through Efficient Approximation

- Closest fit approximation of m-adically represented data by m-1 – adically represented data; repeat for m-2, …
- Computationally this is linear in the number of observations multiplied by the number of digits of precision.
- F. Murtagh, “Sparse p-adic data coding for computationally efficient and effective big data analytics”, p-Adic Numbers, Ultrametric Analysis and Applications, 8 (3), 236--249, 2016.

Spectroscopic redshifts. Initial m-adic display, for m = 10. Three digits of precision used.

(Repeated in the next slide.)



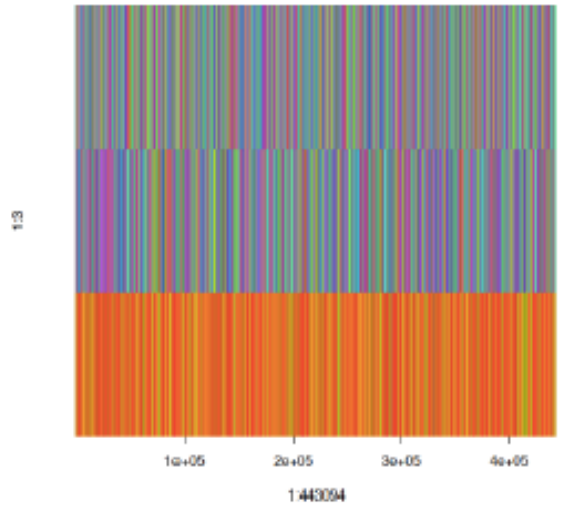


Figure 5: Spectroscopic redshifts. Initial m-adic display, for m = 10. Three digits of precision used.

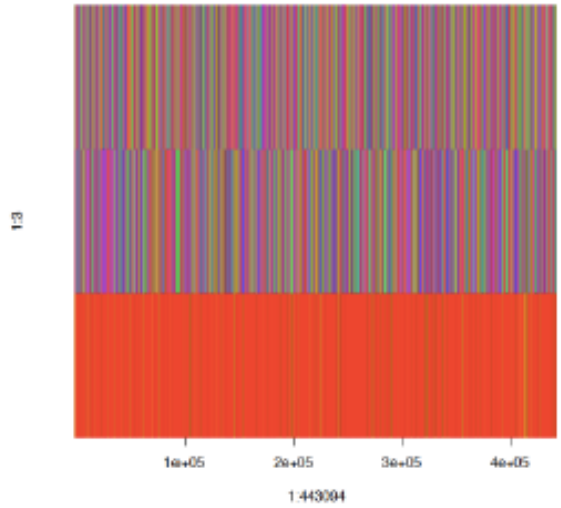


Figure 6: Spectroscopic redshifts. m-Adic display, for m = 9.

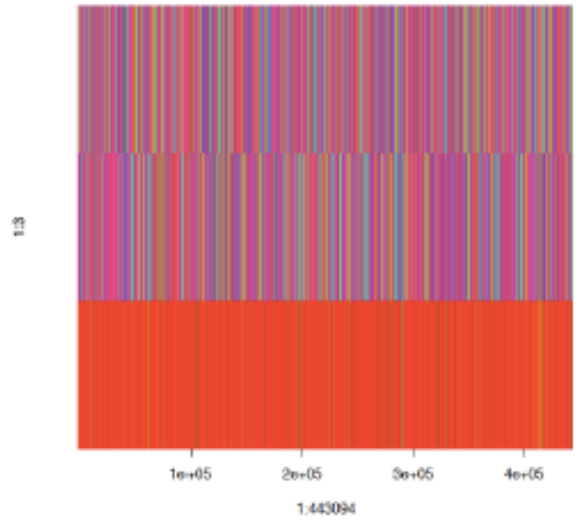


Figure 7: Spectroscopic redshifts. m-Adic display, for $m = 8$.

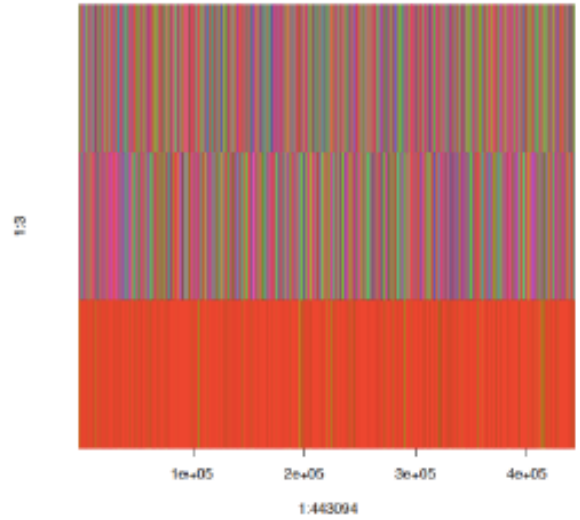


Figure 8: Spectroscopic redshifts. p-Adic display, for $p = 7$.

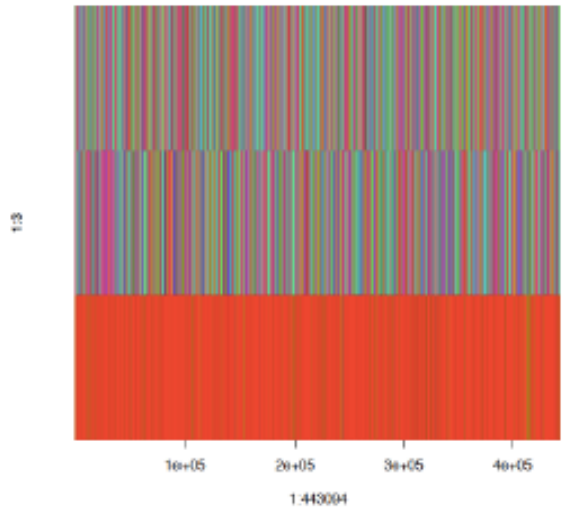


Figure 9: Spectroscopic redshifts. m-Adic display, for $m = 6$.

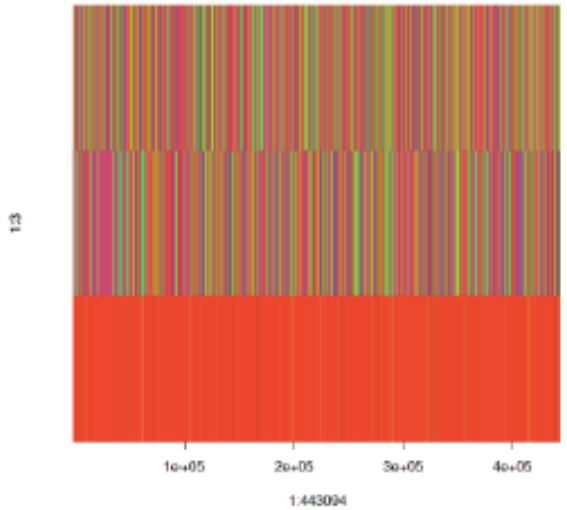


Figure 10: Spectroscopic redshifts. p-Adic display, for $p = 5$.

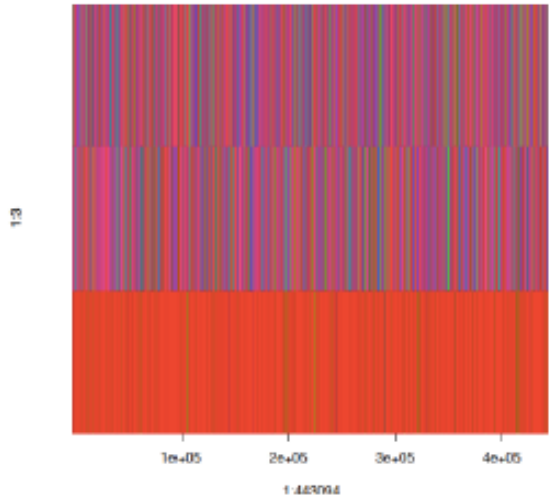


Figure 11: Spectroscopic redshifts. m-Adic display, for $m = 4$.

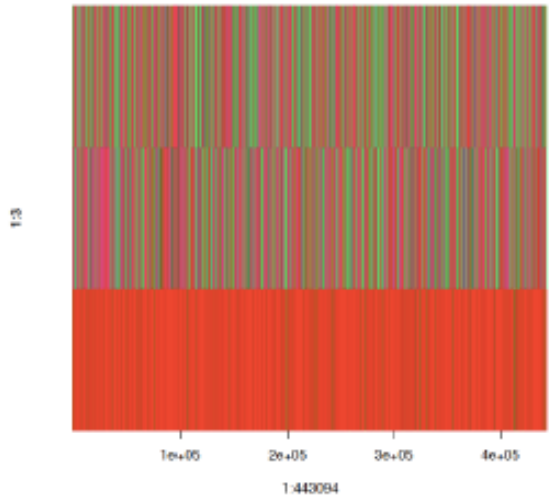


Figure 12: Spectroscopic redshifts. p-Adic display, or ternary, for $p = 3$.

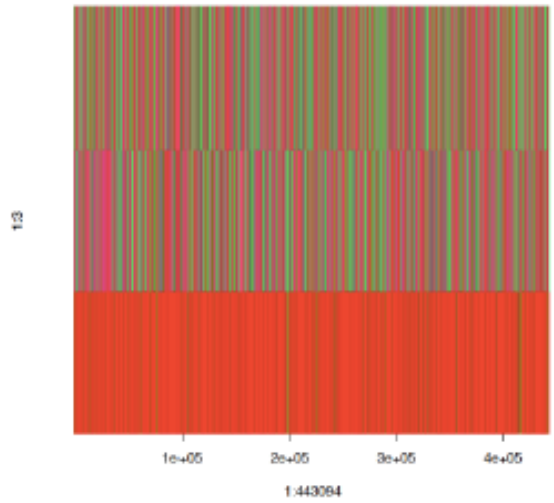


Figure 13: Spectroscopic redshifts. p-Adic display, or binary, for $p = 2$.

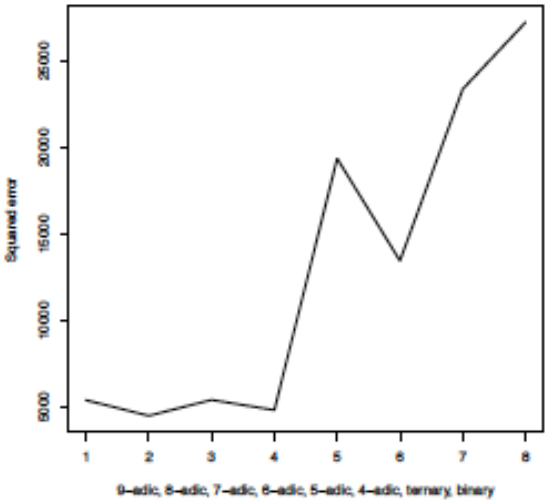


Figure 14: Spectroscopic redshifts. Squared distance, i.e. error, original 10-adic representation, and the sequence of m-adic best fits.

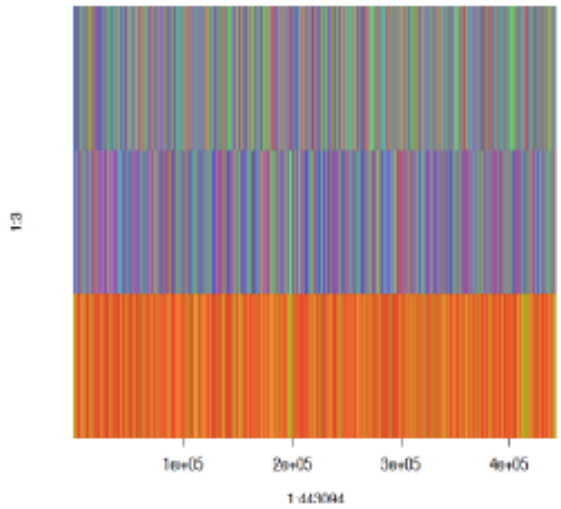


Figure 15: Photometric redshifts. Initial m-adic display, for m = 10. Three digits of precision used.

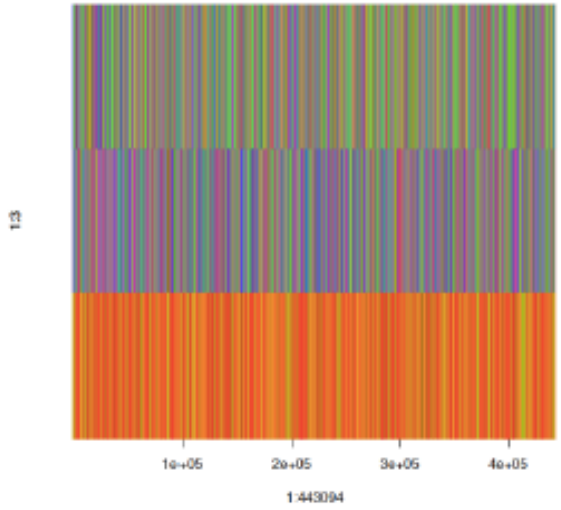


Figure 16: Photometric redshifts. m-Adic display, for m = 9.

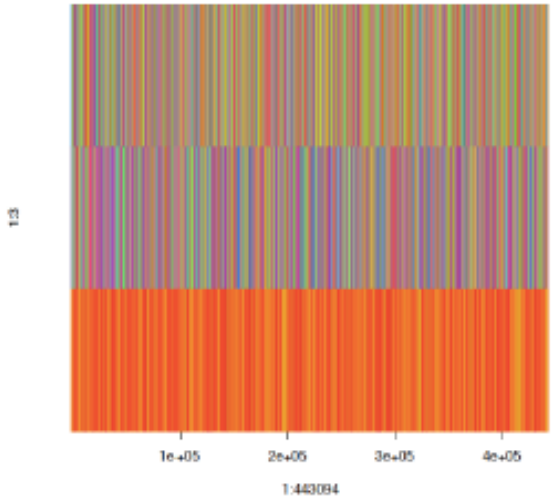


Figure 17: Photometric redshifts. m-Adic display, for $m = 8$.

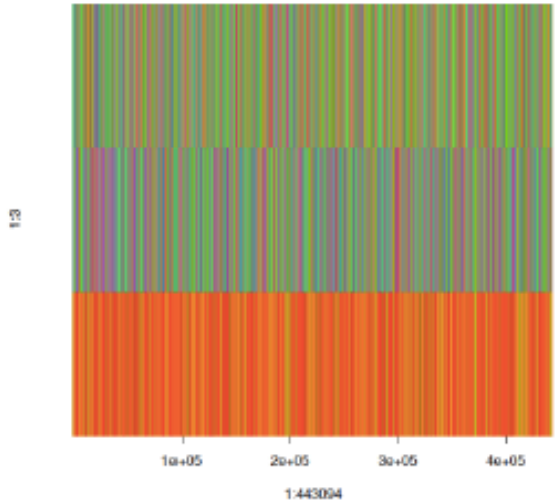


Figure 18: Photometric redshifts. p-Adic display, for $p = 7$.

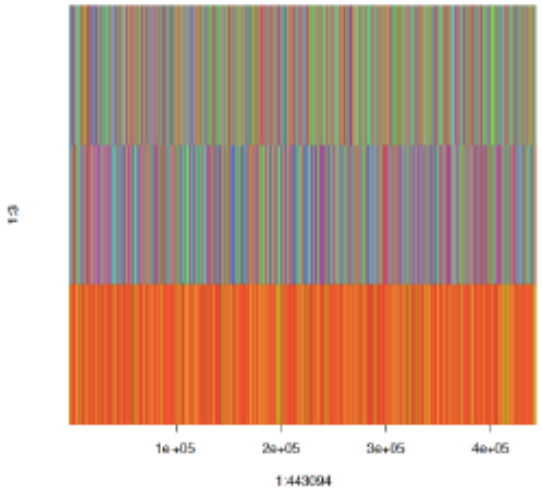


Figure 19: Photometric redshifts. m-Adic display, for m = 6.

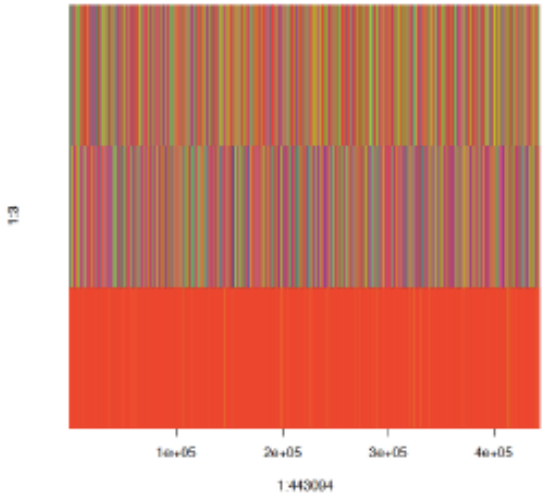


Figure 20: Photometric redshifts. p-Adic display, for p = 5.

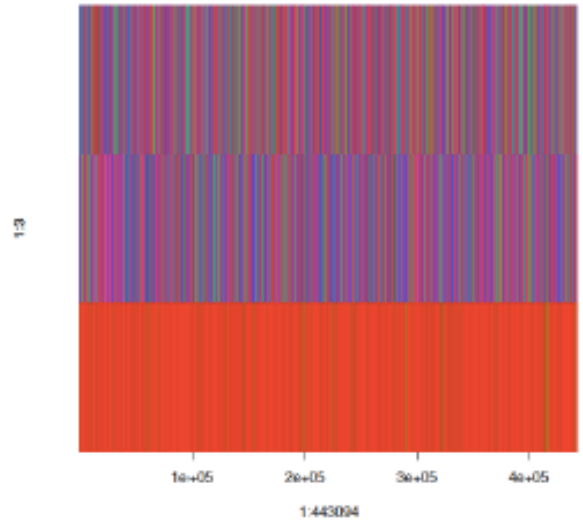


Figure 21: Photometric redshifts. m-Adic display, for $m = 4$.

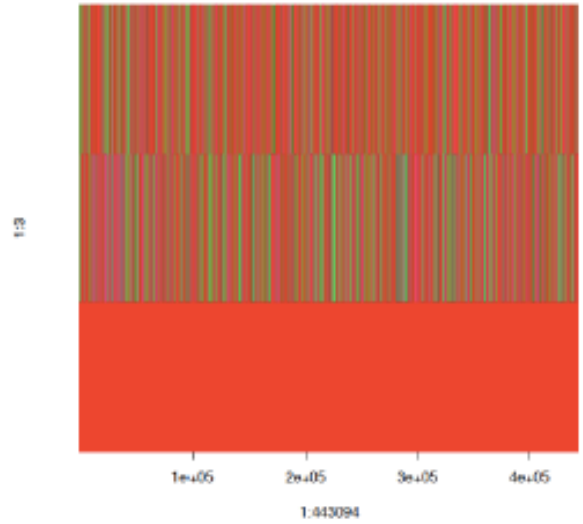


Figure 22: Photometric redshifts. p-Adic display, or ternary, for $p = 3$.

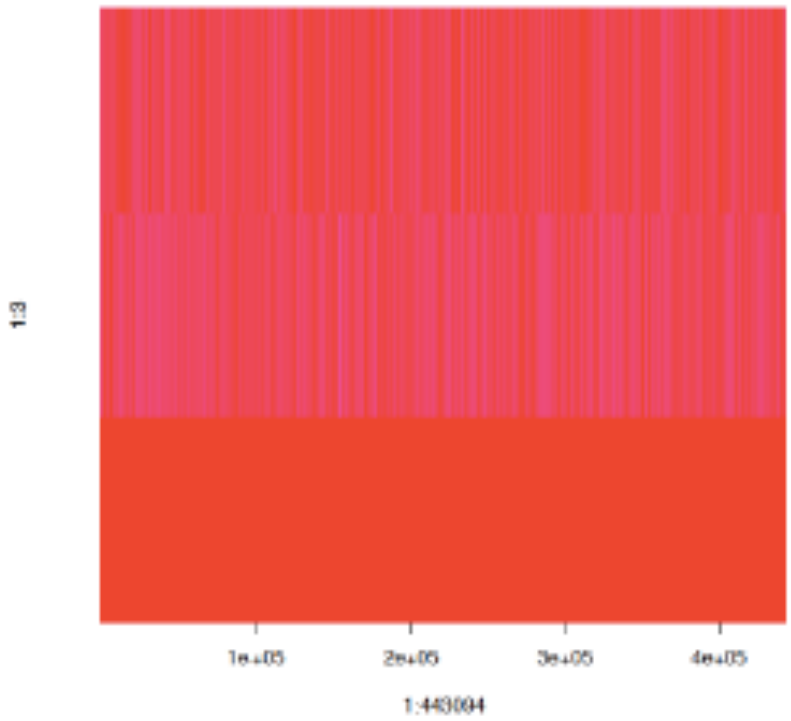


Figure 23: Photometric redshifts. p-Adic display, or binary, for $p = 2$.

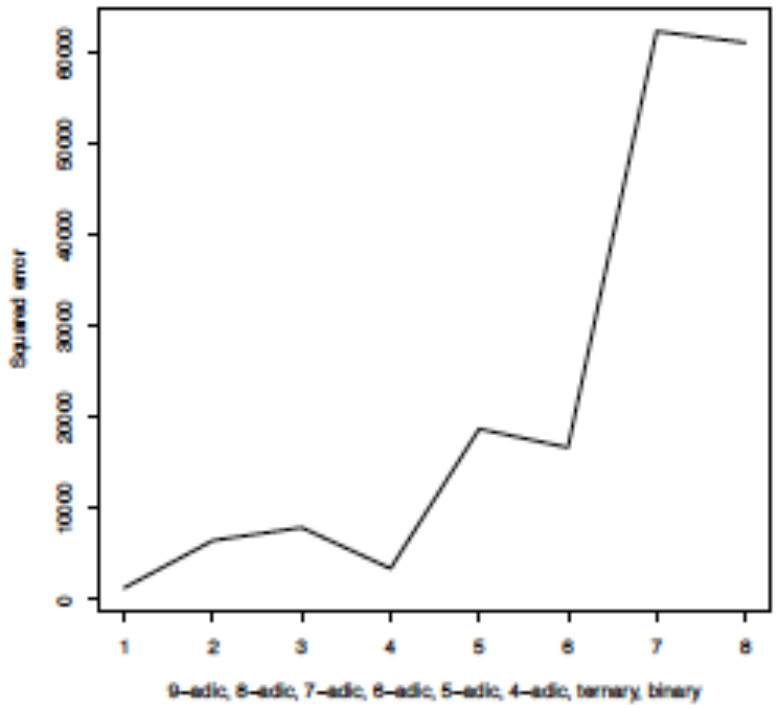


Figure 24: Photometric redshifts. Squared distance, i.e. error, original 10-adic representation, and the sequence of m-adic best fits.

Totalled distances between spectroscopic and photometric redshifts

Representation	Distance
Original, m-adic	3497.347
m-adic, m = 9	3219.545
m-adic, m = 8	2960.628
p-adic, p = 7	2463.237
m-adic, m = 6	2102.937
p-adic, p = 5	1798.283
m-adic, m = 4	1401.31
p-adic, p = 3	1009.443
p-adic, p = 2	940.8114

Table 1: Totalled distance between

- In Table 1, the binary representation of the spectroscopic and photometric redshifts gives the best, closest correspondence.
- In Table 2, to follow, up to 57% of the digits in the ternary, 3-adic, representations of the spectroscopic and photometric redshifts are identical. Next, improve this.
- Table 3: the first digit of the representation of the redshift values is used. For either p-adic with p = 5, or m-adic with m = 4, we have 98% identity between spectroscopic and photometric redshifts.
- Thus: desirability of either 4-adic or 5-adic redshift encoding. I.e. values using digit sets 0,1,2,3 or 0,1,2,3,4.
- In all number theory representations, there is a natural, implicit hierarchical data represenation.
- For p-adic with p = 2, p = 3, i.e. binary, ternary representations, spectroscopic and photometric redshift identity is just over 89%.

Identical digits between spectroscopic and photometric redshifts, the total number, and as the fraction of all digits in these 443094 objects

Representation	No. identical digits	Fraction
Original, m-adic	508376	0.3824441
m-adic, m = 9	361332	0.2718249
m-adic, m = 8	404957	0.3046434
p-adic, p = 7	446470	0.3358731
m-adic, m = 6	487841	0.3669959
p-adic, p = 5	712084	0.5356907
m-adic, m = 4	745784	0.5610427
p-adic, p = 3	757357	0.5697489
p-adic, p = 2	736578	0.5541172

Table 2: Identical digits between spectroscopic and pl
total number, and as the fraction of all digits in these

Compared to previous table, here just the first digit of precision is used. Identical digits between spectroscopic and photometric redshifts, the total number, and as the fraction of all digits in these 443094 objects

Representation	No. identical digits	Fraction
Original, m-adic	366907	0.8280568
m-adic, m = 9	213872	0.4826786
m-adic, m = 8	247360	0.5582563
p-adic, p = 7	262474	0.5923664
m-adic, m = 6	262474	0.5923664
p-adic, p = 5	434736	0.9811372
m-adic, m = 4	434736	0.9811372
p-adic, p = 3	395490	0.8925646
p-adic, p = 2	395490	0.8925646

Table 3: Compared to Table 2, here just the first digit of
tical digits between spectroscopic and photometric redsh
and as the fraction of all digits in these 443094 objects.

Conclusions

- Acknowledged that the training set used here only.
- This is a clusterwise regression, generalizing nearest neighbour regression.
- Precision of measurement is fundamental. Our focus has been on clustering, or binning, or interval specification.
- Longer term, our objective includes inferring structure from data. Such structure includes relative distance from the observer, and associated with this, inter- and intra-distances for clustered objects. Central to this is topology (spatial, shaped, ordered data) rather than geometry. Another longer term goal is the explicit incorporation of the time dimension.

- F. Murtagh, “On ultrametricity, data coding, and computation”, **Journal of Classification**, 21, 167-184, 2004.
- F. Murtagh, “Identifying the ultrametricity of time series”, **European Physical Journal B**, 43, 573-579, 2005.
- F. Murtagh, G. Downs and P. Contreras, “Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding”. **SIAM Jnl. on Scientific Computing**, Vol. 30, No. 2, pp. 707-730. February 2008.
- P. Contreras and F. Murtagh. "Fast, linear time hierarchical clustering using the Baire metric". **Journal of Classification**, 29, 118-143, 2012.
- F. Murtagh and P. Contreras, "Fast, linear time, m-adic hierarchical clustering for search and retrieval using the Baire metric, with linkages to generalized ultrametrics, hashing, formal concept analysis, and precision of data measurement", **p-Adic Numbers, Ultrametric Analysis and Applications**, 4, 45-56, 2012.
- P. Contreras and F. Murtagh, “Linear time Baire hierarchical clustering for enterprise information retrieval”, **International Journal of Software and Informatics**, 6, 363-380, 2012.

- F. Murtagh, “Hierarchical trees in n-body simulations: relations with cluster analysis methods", **Computer Physics Communications**, 52, 15-18, 1988.

- “Clearly those interested in (re)structuring data for any purpose ought to keep a close watch for innovative and interesting approaches in the cosmological simulation field in the future!''. This continues: ``However, the structuring of particles with a view towards force calculations has also something to learn from experience in the cluster analysis field.''

- F. Murtagh, M. Spagat and J.A. Restrepo, “Ultrametric wavelet regression of multivariate time series: application to Colombian conflict analysis", **IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans**, 41, 254-263, 2011.