Response document

Sayani Gupta 28 February 2022

I thank my examiners for their thorough and constructive comments. The point by point description of changes are below: the examiners' comments are in red and my response is in black.

1 Professor Catherine Hurley

1.1 Chapter 1

This would be a good place to detail the co-authorship arrangements for the succeeding chapters, and for the student's contribution to be stated clearly.

A table outlining the main chapters' co-authorship arrangements included in Introduction on page 5.

1.2 Chapter 2

Q1. In Figure 2.1 on the right hand side it should be $t/24$, $t/(24 \times 7)$

Done.

Q2. Last sentence of Section 2.3.4: fix the table reference

Done.

Q3. On page 23 the references to parts of Figure 2.4 need fixing.

Reference is in an acceptable format. No change as already published.

Q4. In Figure 2.4, the outlier glyphs are in my opinion too big and dark and attract too much attention. The log scale for the y-axis should be mentioned in the caption.

The outliers are made smaller and transparent and caption changed.

Q5. The violin plot in Figure 2.5c is barely recognisable, and comparison of the distributions is challenging. The plot deserves a bit more space.

Not changed as figure is already published. Furthermore, the figure already occupies one page.

1.3 Chapter 3

Q1. On page 33 there is a mention of the hakear package, without a reference.

No change. It appears in the abstract of the paper. For the thesis, I have added it since the abstract acts as a summary to the chapter. The package reference is mentioned in the Supplementary section of the paper. Moreover, the plan is to bring all the functions from `gravitas` to `hakear` when we publish it.

Q2. In the Introduction state the methods are for continuous variables.

Added "for a continuous univariate dependent variable" in the last paragraph of Introduction of Chapter 3 on page 39.

Q3. In Figure 3.1, the points are a bit hard to see, so it looks like whiskers extend beyond the data.

Done. Changed to theme_bw() instead of default and outlier.alpha (0.5 ->0.2) and geom_jitter(alpha = 0.04 -> 0.03). Also jitter-width decreased to 0.2.

Q4. On page 35, refer to months consistently using short or long names.

Done. x-axis labels changed to month-of-year for Figure 3.1 and 3.2.

Q5. On page 35 v refers to the number of variables, in Section 3.3.3 it refers to the variable.

Done. On p35, v changed to p, where p represents the number of variables. This is analogous to many time granularities and their pairwise combinations and v represents the univariate measured variable for which distributions are constructed for different time granularities and their pairs.

Q6. In the Figure 3.2 caption, I do not understand the comment beginning "Difference between the 90th. . .". Also for the sentence "Energy consumption for (a)..", it would clarify to insert the word "median".

Done. Changed to "distribution of energy usage"

Q7. On the top of page 38, it should say "the distribution means are three standard deviations apart".

Done.

Q8. Maybe include a reference for gestalt theory.

Done.

Q9. In Section 3.2.2, I find the references to the null distribution confusing. Surely that is the design in Figure 3.3(a) only?

Added text in the caption of Figure 3.3(a) and Section 3.2.2 (Notation) on page 41 to make it clear.

Q10. In table 3.1 N_c is the number of cyclic granularities, whereas in the text (first sentence of 3.3.2 and 3.3.3) it is m.

Added m to table 3.1
m: number of cyclic granularities to display together

Clarification: N_c refers to the total number of contextual cyclic granularities. m refers to the number of cyclic granularities we are considering together in the display. For example, contextual cyclic granularities could be $hour_{day}$, day_{week} and $month_{year}$ and we want to visualize any one granularity at a time. So $N_c = 3$ and $m = 1$.

Q11. Add v to table 3.1.

Added v:continuous univariate measured variable to the table on page 42.

Q12. In the first sentence on the top of page 42, ordered and unordered are mixed up. Have you considered the setting where the facet variable levels are ordered?

Done.

For the within-facet ordered distances, you could consider a distance measure that respects circular order, or choose the start level appropriate to the display.

This is a good suggestion. We will include that into future work.

Q13. I like Figure 3.4. In (3), the dotted arcs only connect to a1 which might be misleading.

Caption changed to address this comment.

Q14. In the pairwise distance measure on page 43, is there any adjustment made for varying numbers of observations across the levels of A and B? For example in Figure 3.3(d) if there are few values at xlevel=1 the comparison shown is less interesting

Added text in the section 3.2.3 under subsection "Characterizing distributions" on page 43 to address this comment.

Q15. The equations on page 44 could be tidied up. There are extra () on the definition of wpd_{perm} and on the residual definition at the bottom of the page. Use $\times$ for the equations.

Done.

Q16. Is there a practical reason why wpd_{glm} is not working as expected for lower nx and nf?

Added the empirical reason in Section 3.2.4 under sub-section "Combination approach" on page 47.

Q17. In the 3.3.1 algorithm, there is m, and then M.

Changed on page 49.

Q18. In Table 3.2, I would suggest separating the m = 1 and m = 2 tables.

$m = 1$ removed from the main chapter and included in supplements.

Q19. The presentation of the material on the simulation study on page 48 could be improved. Where is the notation $wpd_{l,s}$ used? Figure 3.5 shows the m = 2 results only. The text alludes to a simulation involving different underlying distributions, but this is not mentioned again.

Rephrased the simulation design and results corresponding to $m = 2$ on page 50. Similar design and results for $m = 1$, although important, are not included in the main chapter but in the supplements. Simulation results involving different underlying distributions are also presented in the supplementary paper.

Q20. In Figure 3.5 the axis tick labels should be smaller. The blue and orange marks are hard to see. Maybe show fewer panels?

Changed (page 51). Axis tick labels made smaller and space between facets minimised to allow for more space.

Q21. In Figure 3.6 the axis tick labels should be smaller. The blue and orange are hard to see. The caption should refer to the rug. Maybe show fewer panels?

Changed (page 52). Axis tick labels made smaller and the rug plots are removed to make the display less busy.

Q22. In Figure 3.7 (a) the heatmaps need id labels. The grey color is missing from the legend. Maybe use a different colour in the heatmap for the significant comparisons. State the threshold for significance in the caption. In Table 3.3 the caption should explain the threshold. (Maybe use colour instead of stars to indicate significance?)

Changed (page 53). - Different box borders (instead of id labels) are added such that (a) and (b) display the household ids in the same order as indicated by the same colour of the line plot in (b) and box border
- threshold explained in caption in Table 3.3
- grey color explained in the text and caption

Clarification: Using color to convey significance instead of stars (in the Fig 3.3a) is a not a great idea because the varied shades of the tiles represent different values of the wpd. For example, both tiles may be significant, and their colours may differ to represent distinct wpd values. To be consistent, Table 3.3 also has stars to represent significance.

Q23. The link in the Acknowledgements https://github.com/Sayani07/paper-hakear is not available. Neither are the supplementary materials. I understand this is for the paper version, I mention it for completeness.

The Github repository has now been made public. The link works now.

1.4 Chapter 4

Q1. In Section 4.1, line 5 "method of time series clustering".

Done.

Q2. Page 59, fourth bullet point. I found these sentences confusing.

Rephrased to make it clear. (page 61)

Q3. In the material in the bottom of page 59, make it clear from the outset you only have data on energy use, not on property size, location, family size and so on.

Added the statement in the first paragraph of page 62.

Q4. End of Chapter 4.1: incorrect reference to Section 2.7.

Done.

Q5. I found Figure 4.1 very useful. In each box, would be helpful to put in "or", in places where only one of the steps listed is performed, eq Normal quantile transform or Robust scaling. As there are many steps in the algorithm, it would be helpful to the reader to label the pipeline steps and to refer back to them in the text. The text in the Data pre-processing box does not make sense to me.

Rephrased the text in the section 4.2.3 (Data pre-processing) to make it easier to understand. Figure 4.1 labels changed to make it consistent with the text.

Q6. The section describing RS and NQT is confusing. It states RS is applied to each time series separately. Is NQT also applied to each observation separately? How does "it could be useful to standardize it for the selected set of significant granularities prior to computing the distances" relate to the following bullet points?

Text added under Section 4.2.2 on pages 64 and 65 to clarify this.

Q7. Page 64 "D is the Jensen-Shannon distance".

No change. I had mentioned that D stands for Jensen-Shannon distances (now page 66).

Q8. In Table 4.1, the R column should have 250, not 20.

Done.

Q9. The description of the data generation on page 67 and Table 4.2 is confusing. Maybe state the distribution of each v_t.

The table format and caption changed to address this comment. (page 70)

Q10. Where is μ in Table 4.2?

The changed caption (Table 4.2 page 70) will help clarify this.

Clarification: μ is the difference between means considered for consecutive categories. It will vary depending on the chosen design and hence not tabulated.

Q11. In Section 4.3.3 what are the values of R and T?

Added in the first line of Section 4.3.3 (page 72.)

Q12. Figure 4.5 caption: MDS summary plots of what?

Added in the caption (page 75) to address this comment.

Q13. Figure 4.7 is missing labels (a), (b) (c).

Added a, b representing 12 customers (rows) in each of them (page 78).

Q14. Why do you chose to summarize 4 and 5 clusters from JS-NQT when sindex suggests 3 clusters?

Caption (Figure 4.9 page 80) and description in the text (page 78) changed to address this comment.

Q15. The reference to the stationarity assumption in the Discussion needs Clarification.

Text changed in the Discussion section (under Section 4.5.2) to address this.

Q16. Page 81 "can not" should be "cannot"

Done.

Q17. Again, the computational burden could do with more Discussion. At present, if I were to use this method on my data, what kind of sizes are realistic to work with?

A paragraph is added to the Discussion section to incorporate comments on the computational burden. (second paragraph of section 4.5.2 on pages 83, 84)

2 Prof Juergen Symanzik

2.1 Chapter 1

Q1. While the author hints at other sources of similar data, I think it would be worthwhile to give the reader some further specific ideas here in the Introduction where similar complex data may occur, e.g., data from traffic sensors, movement data (turnstiles in metro stations or public buildings), and even more traditional measurements such as temperatures or precipitation over time that could benefit from the methods and software described in this thesis.

Added in the Introduction (page 2).

Q2. It may also be helpful to state what is not be covered in this thesis: The spatial component of such data where an additional component could be latitude and longitude or some areal code, such as the post code or some administrative units. Extending the work from this thesis with spatial proximity methods for clustering (https://doi.org/10.1016/j.cageo.2011.12.017) [assuming that households in certain residential neighborhoods may have similar energy usage patterns] and trying spatial visualization methods (if possible) such as glyph maps (https: //onlinelibrary.wiley.com/doi/abs/10.1002/env.2152) should be mentioned somewhere as a consideration for future work but that may become a future new thesis by itself.

The name of the thesis implies that only analyzing temporal data is covered. Extending it to spatial is too broad a scope. The limitations of the study cater to aspects that could be addressed with some more time. This is not the case for extending the spatial component, which, as accurately pointed out, would require time comparable to another thesis.

Q3. Finally, it would be helpful to indicate here (and in each chapter) what the names or abbreviations for the three R packages (gravitas, hakear, and gracsr) represent. Even the full R package title of "gravitas: Explore Probability Distributions for Bivariate Temporal Granularities" does not provide a full answer what the "tas" component means (I am assuming that Gravi stands for Granularities Visualization). I am even more lost with hakear.

Added under Thesis structure in Introduction on page 4. gravitas: GRAnularity VIsualization for Time-series AnalySis hakear: HArmonies KEeper And Rater gracsr: GRAnularity CluStering in R

2.2 Chapter 2

Q1. One reference to a table (Table 2.3) does not resolve and appears as **ref**{tab:tab-mayan} in the main text on p. 16.

Done.

Q2. how to define and handle pay day (or loan day) which often is defined as the 1st (and/or possibly 15th) work day of a month, resulting in some aperiodic granularity as some of these days across the year could initially fall on a weekend, but even more in a few cases, may coincide with a public holiday (and thus shifting pay day by 2 or 3 days). Similarly, extended weekends that could stretch from Fridays to Sundays or from Saturdays to Mondays (and possibly even over a 4{day period) might be worth some discussion, in particular as a case of variable{length aperiodic granularity. This latter granularity may be of particular interest when working with temporal data with an economic (travel, restaurants, etc.) or entertainment (movie theaters, casinos, etc.) aspect.

Very good point. Paper is already published. We would incorporate this in our future work.

2.3 Chapter 3

Q1. In Figure 3.1, it would help the reader to mention in the figure caption that a log-scale has been used for the vertical axis. Moreover, it might be helpful to add a ticmark label at 10 kWh and add ticmarks representing each fraction of 10 on the log-scale and not only the mid-point between two ticmark labels. Also, reminding the reader that summer months are in January and February in Australia would be helpful. When looking at the figure (prior to reading the main text), my first interpretation for the increased energy usage in January and February was because of (electric) heating and not cooling.

Done.

Q2. Readers may not be familiar with administrative units in Australia. So, instead of speaking of Victoria on p. 35, simply speak of Melbourne (as this seems to be the source of the data).

Done.

Q3. In Figure 3.2, two households are being compared. It would be much easier for most readers if the graphs use a common scale, thus following the small multiple principle. Both vertical axes should be extend to 1.5 kWh and a ticmark label should be placed there as well.

Done.

Q4. In general, I like to see ticmark labels close to the extrema of the shown data points. So, in Figure 3.3a, there should be additional ticmark labels at -3 and +3, in 3.3b at -5 and +10, and so on.

Done.

Q5. Before introducing a new distance measure in Section 3.2, I would like to see a literature review of existing distance measures and their limitations. Why is it necessary to introduce a new distance measure here? This is partially addressed on p. 41, but that should be placed earlier in the text. One clarifying question: Which density is represented by f in the equation? And shouldn't there be a dx at the end of the integral? Also, the distance measures that are mentioned at the end of that paragraph should be explained in one or two sentences.

Changed the following:

- Introduction (last but one paragraph on page 39) to incorporate comments on existing distance measures and their limitations
- "Distance between distributions" (under section 3.2.3 on page 44) to mention why we chose we chose Jensen-Shannon distances and some details about other distance measures.
- f in the equation was a typo. Replaced by log now.

Q6. For the data transformation steps on p. 40, a small table, say with a sample of 5 values from an exponential distribution, might be helpful to better explain each of the three steps.

We will include this in the supplements while publishing the paper.

Q7. A few editorial corrections are necessary, e.g., Dang and Wilkinson (2014); Wilkinson, Anand, and Grossman (2005) provide misses an \and" between the two references. The same holds for Buja et al. (2009); Majumder, Hofmann, and Cook (2013) present. I would leave it to the author to check for similar omissions. Moreover, these should be past tense: "provided" and "presented".

Made it past tense.

Q8. The term Gestalt theory is mentioned a few times (p. 38 & p. 42), but it is never supported by a reference.

Done.

Q9. It would be helpful to add a specific link to a subsection in a cross-reference such as (See the supplements for more details.) on p. 42. Same on p. 47 and p. 48.

Mentioned which table or subsection needs to be referred and the link of the supplementary material provided.

Q10. On p. 45 and in (3.2), it is not immediately clear whether nx; nf simultaneously have to be less than or equal to 5 so that wpd_{perm} is being used - or whether only one of them has to be less than 5. Just a verbal clarification is needed. A practical scenario where this applies might be months (12) and weekdays/weekend (2).

Clarification: We have assumed $nx; nf$ simultaneously to be less than or equal to 5

Q11. On p. 47, Again consider 3.1(a) and 3.1(b) seems to miss the word "Figure".

Done.

Q12. In Table 3.2, listing a p-value as 0 never is a good idea. List it as < 0.01 or < 0.0001 or any other meaningful threshold that matches the number of your simulation runs.

Done.

Q13. The Results section and Figure 3.5 need some clarifications. The text states: Figure 3.5 shows that both the location and scale of the distributions change across panels. I suppose this figure relates to m = 2, but this is not stated in the main text. Table 3.2 lists both, m = 1 and m = 2. Moreover, in Figure 3.5, the y-axis ticmark labels overplot and the nx values are partially cut off.

Done. We decided to remove the y-axis labels in the interest of space. The y-axis is showing the density values but we are only interested in its shape.

Clarification: Added $m = 2$ in the caption of Figures 3.5 and 3.6.

Q14. There are similar problems with the labels in Figure 3.6. Moreover, the two overlaid curves are very hard to distinguish. Would it help to change colors and transparency, or draw the boundary of the curves with the specific colors (and not in black)? This may require some experimentation. Finally, using a differently colored background (say light yellow) would be helpful to visually demonstrate where wpd_{perm} is being used. For the legend, use the terms as in equation (3.2) and not the full word.

Done. Changed to make it more clear than earlier.

Q15. The last paragraph on p. 48 uses $<$ while equation (3.2) uses $\leqslant$. Which one is correct?

Done. $\leqslant$ for both.

Q16. Match the text on p. 50 and introduce the abbreviations from Figure 3.7 and Table 3.3, e.g., "hod" likely is "hour day" and so on. Introducing these abbreviations on p. 53 comes too late.

Added the abbreviations on page 54 in "Choosing cyclic granularities of interest and removing clashes" under section 3.4.

Q17. In Figure 3.7a, it is not clear what the eight heatmaps represent. Do they belong to id 1, . . . , id 8 ? Some explanation in the text or listing the id information on the left would be helpful. Similarly, on p. 53, the text states: id 7 and 8 have the same significant harmonies. I do not know where to look at the heatmaps in Figure 3.7a to find this information (and what is mentioned under items 2. and 3. in the text). Some explanation and labeling is needed here.

Caption changed on page 53 to illustrate the plot better. Added box borders in (a) such that (a) and (b) display the household ids in the same order as indicated by the same color of the line plot in (b) and box border in (a).

Q18. How were the harmony pairs selected and sorted in Figure 3.7 and Table 3.3? headmaps show 25 pairs, the parallel coordinate plot shows only 14 (same as Table 3.3). However, the last two are sorted differently, "hod wdwnd" (the first row in the table) is neither the first nor the last row in the parallel coordinate plot.

Clarification 1: Each household is represented by 25 tiles, each tile representing a pair of cyclic granularities. The colors (in shades of red) represent the value of wpd for each of the harmony pairs (in Table Table 3.3) and the grey tiles correspond to clashes. A darker shade of red corresponds to higher values of wpd.

Clarification 2: The harmony pairs in the parallel coordinate plot are arranged from highest to lowest values of wpd averaged over all households and hence do not correspond to the first or last rows of Table 3.3.

Changed the text description of Figure 3.7 on page 55.

Q19. On p. 53, you use the term inconsequential.This hasn't been introduced. Do you mean "not significant" here?

Changed to "insignificant".

Q20. Apparently, Figure 3.8 uses a logscale again. Adjust similar to my comments for Figure 3.1.

Done.

Q21. Are parts a and b in Figure 3.8 correctly labeled and also matching with the figure caption? The caption reads For id 1, patterns look similar within weekdays and weekends. Visually, this seems to be the case for id 7 (in part b). This also visually matches Figure 3.7b.

Caption rephrased to make it more clear on page 56.

Q22. I didn't ask earlier, but how were the eight households for Section 3.4 being selected? Was this based on a random sample, or manually do obtain rather different energy usage patterns? This should be answered earlier on, but this also should be addressed in the Discussion: What can be expected when working with the full data set of Chapter 2, i.e., about 13,000. You indicate: A future direction of work is to be able to explore and compare many individuals/subjects together for similar patterns across significant granularities. What would be computationally (and visually) feasible as 13,000 is a few orders of magnitude higher than the number of eight households used in this example? Clearly, no full answer is expected here, but rather some speculation of what could possibly be done in the future (or what is done in the next chapter).

- Text added in the first paragraph of Application to elaborate how households were selected (page 51).
- A paragraph is added to the Discussion section to incorporate comments on computational time (page 57).

Q23.The Supplementary materials section mentions data and scripts and a supplementary paper. For the thesis, it would be helpful to indicate where these can be located on github. There is only a link for the R package.

Done.

2.4 Chapter 4

Q1. There seems to be some contradiction on p. 60. One sentence states: Tureczek and Nielsen (2017) conducted a systematic study of over 2100 peer-reviewed papers on smart meter data analytics. Another sentence states: Time series data, such as smart meter data, are not well-suited to any of the techniques mentioned in Tureczek and Nielsen (2017).

Rephrased Section 4.1 on page 62 to make it clear that most techniques mentioned in Tureczek and Nielsen (2017) do not treat smart meter data as a data type with temporal component.

Q2. Use a consistent style and use past tense throughout when you summarize what was previously published.

Done.

Q3. The last cross-reference at the end of the Introduction should be to Section 4.5 (and not Section 2.7).

Done.

Q4. At the end of p. 61, the sentence The flow of the procedures is illustrated in Figure 4.1. should be extended with "and is further described in the following subsections." Also match subsection headings and headers in Figure 4.1, e.g., Selecting granularities vs. Find significant granularities and more.

Done.

Q5. The text on p. 68 speaks of Figure 4.3(b) and Figure 4.3 (right). Either add letters a and b or speak of left and right.

Done.

Q6. Similar to Chapter 3, supplementary material should be specified in more detail, e.g., on p. 68, p. 70, p. 71, and p. 75.

Added links in the supplementary materials section at the end of the chapter.

Q7. In Figure 4.5, it is hard to see which groups are overplotting, in particular for S3. To better reveal this graphically, use different (open) glyphs in addition to different colors. In particular, a + for group 1 and an open circle for group 5 may help to better distinguish the groups (and possibly $\times$, open triangle, and open box to be used for groups 2 to 4).

Done.

Q8. Apparently, in Figure 4.6, the vertical axis is not standardized to [0; 1] as it is frequently done for parallel coordinate plots. Therefore, it would make sense to display an actual vertical axis in each of the three graphs.

No change. Showing the vertical axis would not give us any more information than it is already giving.

Q9. Figure 4.7 speaks of (a) and (b) in the caption, but those letters do not appear in the figure. Either add them or refer to left and right. Moreover, in previous chapters, you used orange and green for the quartiles and 10th/90th quantile. Why not using the same colors and quantiles here? If this becomes too confusing for a reader, then at least, indicate in the caption which quantiles are covered by the gray areas.

Done.

Q10. Be consistent across chapters. For example, in Table 3.3, you use wdwnd. On p. 72 and in Figure 4.7, you use wnwd. Adjust across all chapters. Check whether other abbreviations need to be standardized as well.

Done. Used "wnwd" through out Chapters 3 AND 4.

Q11. Similar to Figure 4.9, it would help the reader that cluster P-3 is visually represented via three somewhat similar colors for Q-3, Q-4, and Q-5. Changing from red to blue/purple initially hides this information. Also arrange the legend from P-1 to P-3 and Q-1 to Q-5 to make it more obvious that P-1 & Q-1 and P-2 & Q-2 are identical.

Done.

Q12. It would help the reader to also mention in the caption of Table 4.3 that P-1 & Q-1 and P-2 & Q-2 are identical. I first looked at Figure 4.9 and then at Table 4.3, but if read in the other order, I likely would have been surprised, to see identical values in some of the table rows (without any further explanation).

Done.

Q13. For the thesis, the work is adequate and meaningful. For a journal paper, I would like to see some of the points outlined in the Discussion being addressed. What happens if this method is applied to the data from the 13,000 customers that were introduced earlier on? As a reader of a journal paper, I would rather like to see limitations of the method proposed here, e.g., if it does not scale up then why it most likely does not scale up. Otherwise, it is hard to assess whether it is worthwhile to try this method for one's own data.

The Discussion section is split into (1) conclusions, and (2) limitations and future research (including a discussion on computational complexity) on pages 82-84 to address this comment.

2.5 Bibliography

Q1. Capitalize all nouns, e.g., in Time-series clustering - A decade review.

Made all bib entries consistent (sentence case)

Q2. Use full{length journal names, i.e., do not use abbreviations, e.g., in Inf. Syst.

Done. Fixed all similar problems.

Q5. Make sure that URLs are clickable and lead to the correct web page, e.g., robjhyndman.com/publications/hakear does not resolve correctly.

Done. All URLs checked and they are clickable.

Q6. For Gupta, S, RJ Hyndman, D Cook, and A Unwin (2021), I would suggest to list the DOI as https://doi.org/10.1080/10618600.2021.1938588 as part of the reference.

Done.

Q7. No need to list the ISBN, e.g., in Xie, Y (2016).

Done.