MONASH University

Thesis Examination Report Summary

Thesis title	Visualization and analysis of probability distributions of large temporal data
Student name	SAYANI GUPTA
Faculty	Faculty of Business and Economics
School	Econometrics and Business Statistics
Course	Doctor of Philosophy
Supervisor name	Rob Hyndman
Overall recommendation	PassWithMajorRevisions
Reports released by the Chair of Examiners on	28/01/2022

Monash Graduate Research Office
Chancellery Building D,26 Sports Walk
Monash University, VIC 3800, Australia
Phone: +61 3 990 54638
Website: www.monash.edu/graduate-research

CRICOS Provider No. 000008C

Examiner: Professor Catherine Hurley

Title	Professor
Name	Catherine Hurley
Current institution	Maynooth University
Recommendation	Pass, with minor amendments

The thesis makes a significant contribution to knowledge and understanding of the field of research.

Thesis Merit

Professor Catherine Hurley

Total number of theses examined	Less than 5
Does the thesis contain material worthy of publication in a form appropriate to the discipline?	Yes
Is the format and literary presentation of the thesis satisfactory?	Yes

The thesis as a whole is a substantial and original contribution to knowledge of the subject with which it deals.	★★★★★ Exceptional
The student shows familiarity with and understanding of the relevant literature.	★★★★★ Exceptional
The research methods adopted are appropriate to the subject matter and are properly applied.	★★★★★ Exceptional
The results are suitably set out and accompanied by adequate exposition.	★★★★★ Exceptional
The quality of academic writing and general presentation are of a standard for publication.	★★★★★ Exceptional

1. **Exceptional** - Of the highest merit and at the forefront of international doctorates in the field. Fewer than 5% of students worldwide would fall in this band.
2. **Excellent** - Strongly competitive at an international level. Fewer than 20% of students would fall in this band.
3. **Very good** - Interesting and sound, Approximately 30% of students would fall in this band.
4. **Good** - Sound but lacking in some respect. Approximately 30% of students would fall in this band.
5. **Fair** - Has potential, but requires major revisions. Approximately 15% of students would fall in this band.

Report on: Visualization and analysis of probability distributions of large temporal data

Catherine Hurley

20/1/2022

Overall

This research is on visualisation techniques for temporal data, where the data is sliced and diced and clustered in various ways in an effort to seek interesting patterns. The main application presented is smart meter data, but the topic is broadly interesting and relevant to wide varieties of time-based data. The techniques are backed up by R software packages which are freely available, making the methods widely accessible.

The treatment of the topic was overall excellent, thought-provoking and is of a very high standard. The thesis was a delight to read and visualisations are well-designed. The work is a significant contribution to exploratory visualisation of time series. I pass on my congratulations to Ms. Gupta on a job well done.

My comments on the thesis are minor.

Chapter 1 Introduction

The thesis is arranged in the format of three separate papers, making up the content of Chapters 2,3 and 4. Chapter 1 gives a general overview of the type of data, and a brief summary of succeeding Chapters. I would have preferred a longer Introduction giving more details on the problem background, even if this material appears elsewhere in the thesis. The Chapter summaries could also be more detailed. But overall this Chapter is good and it does not require any revision. Perhaps though, 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.

Chapter 2 Visualizing probability distributions across bivariate cyclic temporal granularities

This Chapter is "to appear" in JCGS and congratulations are due to Ms. Gupta on this achievement. It sets out definitions for a calender algebra for nested time categories. The material on circular granularities is new and some of the terminology developed here is used elsewhere in the thesis. Notation is clearly defined.

The cricket example is interesting, though somewhat challenging to get my head around as I'm not familiar with the game.

I'm happy to see the package gravitas is complete with vignettes and available to try. I would have liked to see a bit more discussion about the package in the thesis. I particularly liked the idea of `gran_advice` and would be interested in a discussion of it, but maybe this is for another paper.

- In Figure 2.1 on the right hand side it should be $\lfloor t/24 \rfloor$, $\lfloor t/(24 \times 7) \rfloor$.
- Last sentence of Section 2.3.4: fix the table reference
- On page 23 the references to parts of Figure 2.4 need fixing.

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

Chapter 3 Detecting distributional differences between temporal granularities for exploratory time series analysis

This chapter provides techniques for sifting through a multitude of displays of a time series to uncover patterns. The ideas here are very relevant as datasets become bigger providing a challenge to manual exploration. The chapter is well-organised and well-written and is of a standard for publication in a good journal.

I like the idea of constructing a distance measure between distributions, to measure the relevance of the comparison shown in a visualisation. I would also like to see a discussion on the computational burden of the methodology. Some of this is adressed in Chapter 5, however.

- On page 33 there is a mention of the hakear package, without a reference.
- In the Introduction state the methods are for continuous variables.
- In Figure 3.1, the points are a bit hard to see, so it looks like whiskers extend beyond the data.
- On page 35, refer to months consistently using short or long names.
- On page 35 v refers to the number of variables, in Section 3.3.3 it refers to the variable.
- 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".
- On the top of page 38, it should say "the distribution means are three standard deviations apart".
- Maybe include a reference for gestalt theory.
- In Section 3.2.2, I find the references to the null distribution confusing. Surelt that is the design in Figure 3.3(a) only?
- 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.
- Add v to table 3.1.
- 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? 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.
- I like Figure 3.4. In (3), the dotted arcs only connect to a_1 which might be misleading.
- 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.
- 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 $\times$ not $*$ for the equations.
- Is there a practical reason why wpd_{glm} is not working as expected for lower n_x and n_f?
- In the 3.3.1 algorithm, there is m, and then M.
- In Table 3.2, I would suggest separating the $m = 1$ and $m = 2$ tables.

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

Chapter 4 Clustering time series based on probability distributions across temporal granularities

This Chapter describes techniques for clustering time series. The problem presented is well-motivated by the smart meter data discussed in previous Chapters. The new methodology is interesting, well-presented and the material is of publication quality. I will be interested to see the companion software package, when it is ready for primetime. The methodology was applied to just 24 smart meter customers. While this is perhaps is not enough to gain insight on data patterns more broadly, it does illustrate the feasibility of the new clustering techniques.

- In Section 4.1, line 5 "method of time series clustering".
- Section 4.1, line 12 "decade review": reword.
- Page 59, fourth bullet point. I found these sentences confusing.
- 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.
- End of Chapter 4.1: incorrect reference to Section 2.7.
- 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.
- 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?
- Page 64 "D is the Jensen-Shannon distance".
- In Table 4.1, the R column should have 250, not 20.
- The description of the data generation on page 67 and Table 4.2 is confusing. Maybe state the distribution of each v_t.
- Where is μ in Table 4.2?
- In Section 4.3.3 what are the values of R and T?
- Figure 4.5 caption: MDS summary plots of what?
- Figure 4.7 is missing labels (a), (b) (c).
- Why do you chose to summarize 4 and 5 clusters from JS-NQT when sindex suggests 3 clusters?
- The reference to the stationarity assumption in the Discussion needs clarification.
- Page 81 "can not" should be "cannot"
- 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?

Chapter 5 Conclusion

The conclusion chapter has a nice overview of the techniques in the three main chapters. Could additional predictors on the customers be added to the clustering? The material on software and computation could be added to the individual chapters.

Examiner: Prof Juergen Symanzik

Title	Prof
Name	Juergen Symanzik
Current institution	Utah State University
Recommendation	Pass, with major revisions, certified by Monash Chair of Examiners

The thesis makes a significant contribution to knowledge and understanding of the field of research.

Thesis Merit

Prof Juergen Symanzik

Total number of theses examined	More than 10
Does the thesis contain material worthy of publication in a form appropriate to the discipline?	Yes
Is the format and literary presentation of the thesis satisfactory?	Yes

The thesis as a whole is a substantial and original contribution to knowledge of the subject with which it deals.	★★★★★	Exceptional
The student shows familiarity with and understanding of the relevant literature.	★★★★☆	Excellent
The research methods adopted are appropriate to the subject matter and are properly applied.	★★★★☆	Excellent
The results are suitably set out and accompanied by adequate exposition.	★★★★☆	Excellent
The quality of academic writing and general presentation are of a standard for publication.	★★★★☆	Excellent

1. **Exceptional** - Of the highest merit and at the forefront of international doctorates in the field. Fewer than 5% of students worldwide would fall in this band.
2. **Excellent** - Strongly competitive at an international level. Fewer than 20% of students would fall in this band.
3. **Very good** - Interesting and sound, Approximately 30% of students would fall in this band.
4. **Good** - Sound but lacking in some respect. Approximately 30% of students would fall in this band.
5. **Fair** - Has potential, but requires major revisions. Approximately 15% of students would fall in this band.

DEPARTMENT OF MATHEMATICS & STATISTICS

Dr. Jürgen Symanzik
Professor
Utah State University
Department of Mathematics & Statistics
Logan, UT 84322–3900
USA

Tel.: (435) 797–0696
e–mail: symanzik@math.usu.edu
Web: http://www.math.usu.edu/~symanzik/

January 25, 2022

Doctorate Examination for SAYANI GUPTA

This document contains a review of the doctoral thesis by SAYANI GUPTA, entitled *"Visualization and analysis of probability distributions of large temporal data"*. This is a thesis submitted for the degree of Doctor of Philosophy at Monash University, Department of Econometrics and Business Statistics, in 2021.

The thesis is structured in five chapters:

Chapter 1 : Introduction

Chapter 2 : Visualizing probability distributions across bivariate cyclic temporal granularities

Chapter 3 : Detecting distributional differences between temporal granularities for exploratory time series analysis

Chapter 4 : Clustering time series based on probability distributions across temporal granularities

Chapter 5 : Conclusion

I will follow this structure in my review over the next few pages.

Chapter 1: Introduction

In this chapter, the author sets the stage for her work: smart meter energy use data, collected at half–hour intervals for about two and a half years (from October 2011 to March 2014) for about 13,000 Australian households. Naturally, such temporal data is challenging and complex to analyze and visualize due to numerous temporal patterns and distinct human behavioral patterns regarding energy use.

The author provides a brief overview how such data have been analyzed in the past via time granularities, in particular linear time granularities (hours, days, weeks, and months) and cyclic temporal granularities (hour–of–the–day, and weekday vs. weekend). The author connects the question of how to analyze such data to the concept of exploratory data analysis (EDA) that emphasizes that it is necessary to analyze data from multiple perspectives to fully understand complex data. This leads to the overall goal of the thesis (as stated by the author) *"which aims to provide a platform for systematically exploring probability distributions induced by these multiple observations to support the discovery of regular patterns or anomalies, as well as the exploration of clusters of behaviors or the summarization of the behavior."*

The author continues with an overview of the three main chapters in her dissertation, including information where to find supporting software and data (`https://github.com/Sayani07/thesis-SG`).

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.

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.

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.

Chapter 2: Visualizing probability distributions across bivariate cyclic temporal granularities

In this chapter, the author discusses the concepts of time deconstructions using linear and cyclic time granularities. Software is being introduced that helps the user to determine whether pairs of granularities can be examined together or not. The author speaks of harmony or clash, respectively. Data visualization software is being introduced to explore periodicities, associations, and anomalies of the time series data, making use of the granularities as additional categorical variables such as weekend or weekday or public holiday or not. The main goal of this chapter is to provide formal mathematical definitions and characterizations of the different time granularities and to introduce visualization methods and supporting software for the probability distributions that can be derived from the aggregated time series data, conditional on one or more of the granularities.

This chapter has been published online on July 26,2021, in the Journal of Computational and Graphical Statistics (JCGS) (`https://doi.org/10.1080/10618600.2021.1938588`) with Rob J Hyndman, Dianne Cook, and Antony Unwin as co–authors. It can be expected to be published in the print issue of the journal later in 2022 or early in 2023. JCGS is owned by the American Statistical Association (ASA) and is published by Taylor & Francis. It has an impact factor of 2.3 (in 2020) and a 5–year impact factor of 3.3 (in 2020). It is widely considered as the flagship journal in the field of statistical computing and statistical graphics with an acceptance rate of only about 21% (as stated on `https://www.tandfonline.com/action/journalInformation?show=journalMetrics&journalCode=ucgs20`). I can only congratulate the authors to get their work accepted in such a prestigious journal. This is even more impressive as the time between first submission (September 25, 2020) and final acceptance (May 31, 2021) took only about eight months (which is very fast for this journal).

In the thesis, I noticed only one actual problem in this chapter. 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.

While the work in this chapter is solid and complete, there are some options for follow–up work, e.g., in some future conference presentation and/or for future extensions of the software. For example, 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.

Chapter 3: Detecting distributional differences between temporal granularities ...

In this chapter, the author follows up on cyclic temporal granularities, discussed in more detail in Chapter 2. In this chapter, the main goal is on detecting distributional differences between temporal granularities. This is done via a new distance metric proposed by the author. As for Chapter 2, a supporting software package has been created.

While this chapter has not been reviewed for a journal yet, I have a few more specific comments that should be addressed prior to a journal submission of this chapter:

- 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.
- 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).
- 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.
- 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.
- 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.
- 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.
- 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".

- The term `Gestalt theory` is mentioned a few times (p. 38 & p. 42), but it is never supported by a reference.
- 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.
- On p. 45 and in (3.2), it is not immediately clear whether n_x, n_f 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).
- On p. 47, `Again consider 3.1(a) and 3.1(b)` seems to miss the word "Figure".
- 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.
- 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 n_x values are partially cut off.
- 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.
- The last paragraph on p. 48 uses $<$ while equation (3.2) uses $\leq$. Which one is correct?
- 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.
- 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.
- How were the harmony pairs selected and sorted in Figure 3.7 and Table 3.3? The

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.

- On p. 53, you use the term `inconsequential`.This hasn't been introduced. Do you mean "not significant" here?
- Apparently, Figure 3.8 uses a log–scale again. Adjust similar to my comments for Figure 3.1.
- 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.
- 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).
- 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.

Chapter 4: Clustering time series based on probability distributions across temporal granularities

In this chapter, the author follows up on clustering time series across temporal granularities, building on concepts and ideas introduced in Chapters 2 and 3. Two ideas for measuring distances between time series are introduced that are based on probability distributions of the underlying time series data. As for Chapters 2 and 3, a supporting software package has been created.

While this chapter has not been reviewed for a journal yet, I have a few more specific comments that should be addressed prior to a journal submission of this chapter:

- 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).`

- Use a consistent style and use past tense throughout when you summarize what was previously published. You wrote `Tureczek and Nielsen (2017) conducted` and `Only one study (Ozawa, Furusato, and Yoshida, 2016) identified` but also `Motlagh, Berry, and ONeil (2019) suggests` and `Chicco and Akilimali (2010) addresses`. The last two should also use past tense. Check for others that also should be changed to past tense.
- The last cross–reference at the end of the Introduction should be to Section 4.5 (and not Section 2.7).
- 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.
- 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.
- Similar to Chapter 3, supplementary material should be specified in more detail, e.g., on p. 68, p. 70, p. 71, and p. 75.
- 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).
- 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.
- 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.
- 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.
- 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.

- 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).
- 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.

Chapter 5: Conclusion

In this chapter the author summarizes the content of the thesis, including original contributions and the developed software, and revisits possible directions for future research.

Overall, this is a good summary of the work being conducted. As far as I can say, there is no need for any changes here.

Bibliography

Personally, I like to see consistency across all entries of the reference list. This is currently not the case. Unless Monash University has a certain style requirement for the Bibliography, I would suggest to follow the JCGS style that was used for the journal version of Chapter 2. Here are some necessary changes:

- Capitalize all nouns, e.g., in
  `Time-series clustering - A decade review`.
- Use full–length journal names, i.e., do not use abbreviations, e.g., in
  `Inf. Syst.`
- Capitalize journal names, e.g., in
  `SIAM journal on scientific and statistical computing`.
- For books, add the city of publication, e.g., in
  `Springer Science & Business Media.`
  Interestingly, this is done in the journal version of the second reference: Aigner, W., Miksch, S., Schumann, H., and Tominski, C. (2011), Visualization of Time-Oriented Data, London: Springer Science & Business Media.
- Make sure that URLs are clickable and lead to the correct web page, e.g.,

robjhyndman.com/publications/hakear does not resolve correctly.

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

Several of these comments apply to more than just one entry in the Bibliography.

Overall Assessment

While there are numerous small changes and suggested corrections for Chapters 3 and 4, there is no major issue with any of these chapters. I can only congratulate the author for her impressive work, both in the written thesis and the accompanying software. With these edits in place, I would expect to see a smooth review process of Chapters 3 and 4 in some high–level statistical journals, similar to the already published Chapter 2. This research has been very well done overall!

In case of any further questions, please contact me via e–mail (symanzik@math.usu.edu) or phone (435 797 0696).

Sincerely,

Jürgen Symanzik