August 6, 2014

John-von-Neumann-weg 9
Building HPZ
Zürich CH-8093
Phone: +41 78 971 72 50
Email: miller.clayton@arch.ethz.ch

Miroslaw J. Skibniewski, Ph.D.
Editor-in-Chief
Journal of Automation in Construction
Glenn L. Martin Hall 1188
University of Maryland
College Park, MD 20742, U.S.A

Dear Dr. Skibniewski:

Please find our responses to the 26 questions and comments made by the 3 referees. Overall, we believe that we have fully addressed each issue individually in a way that satisfies the requirements of the journal review process.

Each reviewer's general comments were responded to and their individual questions and comments were addressed in detail in this document. The responses are shaded blue following each reviewer's comments and questions. Included also in this submission are two revised manuscript documents: one in which the original manuscript is marked up with the changes and one that is a final draft version with all changes and additions incorporated.

Please contact me with any further comments or concerns regarding this draft.

Kind Regards,

Clayton Miller

Review Comment Responses to Automation in Construction Submission

Paper Title: Automated daily pattern filtering of measured building performance data

Clayton Miller, Zoltán Nagy, Arno Schlueter

July 29, 2014

1 Review 1

1.1 General comments from reviewer

This paper presents a new tool 'Day Filter', an analysis process for analyzing the sensor data generated from BMS. The motive of the research is good, and is useful for effectively analyzing the huge amount of data generated by the Energy Management Systems. Two case studies were also presented to analyze cooling energy and electrical energy from two buildings respectively.

1.2 Response to general comments

We appreciate your comment that the motivation for the research is valid – this is something that we spent a lot of time investigating and hoping to find a new angle in which automation could be achieved in a way that helps finding structure 'real-world' datasets from real BMS systems. This motivation comes from years of effort in trying to utilize these datasets with a lot of issues related to data modeling, etc.

We appreciate the questions focused on clarifying the background and discussing the limitations of the approach. These are useful points that will be very interesting for a reader of the paper.

1.3 Revision items

a) Literature Review included background and related works nicely. However, the authors can further enhance the section which describes the need of a software like this. Page 4 and 5 explains this. But, we feel that little more stress needs to be given in explaining the difficulties faced presently in analyzing the data

We have further developed the last paragraph before Section 1.1 which explains the comprehensive studies completed by the California Commissioning Collaborative in 2010 focused on the adoption of Building Performance Tracking Tools and techniques. This section discusses their findings which emphasize the need for processes that save analysts time in characterization of building performance and implementation of tools.

Additionally, we have provided more details of various conventional automated fault detection and diagnostics (AFDD) techniques to give a broader overview of their capabilities as compared to the *DayFilter* process.

We have added a subsection on building performance simulation calibration, a subfield of performance analysis which relies and transformation of high frequency time-series data into input assumptions for simulation models. This section defines the gap in data preparation techniques used by the building industry for the calibration process and how *DayFilter* can enhance those efforts. We have added a description of the *day-typing* techniques and diversity factor schedule processes that have been developed over the years. The *DayFilter* improves upon these techniques through automated detection of anomalous daily profiles.

b) Page 10 and 11: Conversion of the building raw data into SAX formation requires to be given more information particularly regarding the string representation. For instance, the method of coining the strings aaaa , abaa , acba etc needs to be explained at least for some cases for all readers to follow the further explanations.

We have added a more detailed explanation of how the strings are created at the end of Section 3.2, parts of Section 3.3, and the captions of Figure 4 and 5. We believe that the simplified two day and two week examples with illustrations should now be basic enough that someone unfamiliar with SAX can understand the method of the string creation. Creation of the augmented suffix tree has been explained in more detail as well as the visualization of the suffix tree as a sankey diagram.

c) Page 14 - Visualization part needs to be explained further with the presented Figure 6. Visualization is the place from which the results are inferred and rather than directly moving to the case studies, there should be more explanation in that part regarding the method of inferring the results from these graphs.

A more detailed explanation of suffix trees, sankey diagrams and their combination is added to Figure 5 and accompanying text in Section 3.3. Section 3.5 was expanded to explain the combination of using heatmaps with the sankey diagrams – what they are and why they are used in this context. A brief overview of how an analyst would utilize these visualizations is discussed also. More details in the case studies are also added to give more insight into how the process uncovered specific issues and insight.

d) Limitations of the model (if any) need to be discussed.

The Discussion section was rewritten to provide a better connection to the literature review as well as discuss some of the limitations that the approach has with regards to the defined objectives of AFDD and model calibration. This includes discussions on data quality, pattern resolution, and utilization of the discords as performance problem indicators.

Additionally, a more detailed discussion of limitations is covered in second paragraph of the conclusion, Section 6. This discussion focuses on the fact that the defined process is univariate.

e) After doing the two case studies, whether any changes were able to make to those buildings based on the results? Whether those changes resulted in any energy reductions? If the authors have this information and if they can answer these questions, it will add more meaning to this research.

More specific content was added to the Case Study 1 description regarding specific investigation of discords in the UWC case study. We discuss a performance problem detected and the actions taken to remedy the issue. In the updated discussion section, we focus on the output of the process and how it can be used in a Qualitative AFDD process to set thresholds, etc.

2 Review 2

2.1 General comments from reviewer

This article provides potential automated data procedures for examining the routine operation profiles of whole building power consumption. Overall, this is a good paper and well written.

2.2 Response to general comments

Thank you for the comment regarding the fact that the paper is good and well written. We appreciate your focus on detailed parts of the process and in helping us clarify the steps in the process that many readers could potentially become stuck on – such as interpretation of the graphics and how to choose the parameters.

2.3 Revision items

f) On page 107, the year of reference No. 10 was published in 2005, not in 2004. Please make sure it.

Thank you for pointing out this obvious error in referencing! This discrepancy has been noted and it has been updated with the correct date, edition, and page numbers. A double check of the other references has been completed.

g) From line 109 to line 110, the authors do not study enough references in analytical-based FDD. The example in [19], it is just one example using the simulation models in the process. However, another method called decoupling-based technique [Li and Bruan, 2003 and Zhao et al, 2014] has been applied in the AFDD application of rooftop units and chillers, respectively. This technique uses virtual sensor techniques based on low-cost manufacturing data and does not require simulation tool. The disadvantage of the method is not convenient for large data systems and is required solid-backgrounds of physical meanings and relations of an applied system. Please make sure your statement. Is it correct?

Thank you for this clarification. It is true that an extensive literature review of AFDD has not been completed. That said, we have added additional text in the Background section to emphasize that the proposed process is not intended as an implementation of FDD, but as an exploratory analysis technique that is useful to support AFDD approaches. It is a discovery process which filters patterns that can be further analyzed and used to create AFDD. Therefore, we simply need to state that AFDD requires a

certain amount of raw data preprocessing for the purposes of automating the implementation and utilization process.

We have added the Zhao et al. 2014 study to the review to show an additional type of FDD technique that could benefit from exploratory analysis.

h) From line 113 to 114, please give a clear problem statement why data-driven area is a major research focus? What are differences between data-driven and previous two methods leading to mainly significant points of view in your study?

We have added content to the Introduction and Background section which seeks to further clarify that the proposed process is a compliment of AFDD and calibrated simulation and not necessarily a competitor. The *date-driven* aspect of the proposed approach is as an exploratory step and not only in the context of traditional process data driven AFDD techniques. We have clarified the cause for motivation in the context of AFDD and simulation model calibration. The newly updated discussion section also has been rewritten to clarify how the process addresses these objectives.

i) On line 135, it is not clear what exact differences are between your application and AFDD in defective equipment. If your application is applied in automatically cleansing building data systems, fault equipment data caused by routine operations are also included in the data system. Have you studied this point if the data of some fault equipment operations occur and are included in building data system? Please clarify the point.

The statement at line 135 has been rewritten to provide more clarity to the point being made – that is that Temporal Data Mining shares features AFDD for buildings and uses different terminology.

I interpret the second part of the comment as addressing the data-driven nature of temporal data mining such that structure is filtered from time series data regardless of whether it is faulty or not. This points is quite valid. If a systematic fault is occurring in the majority of the dataset then that pattern will be filtered as a motif candidate instead of a discord candidate. The algorithm at this point will only characterize the performance and not diagnose whether it is good or bad.

The relationship between the proposed process and AFDD has been further clarified in the Discussion section of the paper

j) On line 165, it is not clear what challenges you has developed. What point does a new process differ from Viztool? You said a lack of tools or processes being similar to Viztool. Do you mean you just develop a similar tool to Viztool for the building application?

VizTool is a very general temporal data mining tool that shares many similarities to the proposed process in terms of functionality. In a way, the *DayFilter* process is an updated version of VizTool but is designed specifically for diurnal performance patterns of building energy data. This application inspires the input parameter settings, the visualization of the daily patterns according to sankey diagram with heat map as opposed to a simple tree, and a very specific discussion of the context, input and output of the process according to two large building performance datasets.

k) From line 260 to 261, could you explain why k-means are suitable than others for your application?

Referenced paper [55] uses a form of k-means (Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns) and finds that Euclidian distance is adequate for this application. This information combined with the fact that k-means is very simple and an widely implemented clustering technique influenced the decision to use it. Further discussion was added in the Clustering subsection to explain this fact. It is not within the scope of this project to validate other techniques.

l) Based on Fig, 4 and 5, you clearly explain how the process perform given by the example. However, it lacks of the explanation for Fig. 6 with the same application. Please explain more.

A more detailed explanation of suffix trees, sankey diagrams and their combination is added to Figure 5 and accompanying text in Section 3.3. Section 3.5 was expanded to explain the combination of using heatmaps with the sankey diagrams – what they are and why they're used in this context. A brief overview of how an analyst would utilize these visualizations is discussed also. More details in the case studies are also added to give more insight into how the process uncovered specific issues and insight.

m) Based on line 306, you do not mention how to appropriately select a threshold for each application.

Each of the 5 steps in Section 3 includes an initial discussion outlining the considerations a user should have when choosing the thresholds for the input parameters. The parameter selection question is addressed explicitly in a new Section 5, Parameter Section Analysis, with multiple experiments example showing the differences between a range of possible inputs of the SAX and clustering processes. Section 5 outlines suggestions for initial settings of the parameters.

n) From line 309 to line 317, can you tabulate all motif and discard candidates in Table for conveniently understanding by readers. How do you obtain or generate the patterns of representative of days in a school?

While we agree that tabulating all the motif and discord profiles in a table would be informative, we believe that this would be overwhelming for the reader. Instead, we propose to chose a few of the discord candidates and do a further, deeper analysis as examples. This has been added as an analysis in Section 3 and as Figure 8.

Additionally, it should be noted that Figure 7 was designed specifically to condense all of the motif and discord profiles in a visualized and readable space. For this reason, we would not like to tabulate the results in a less expressive way.

As for the last question, we assume that the reviewer is inquiring how the clustering patterns were generated in Figure 9. These profiles are an aggregation of the motif candidates using a k-means clustering algorithm. This process is completely automated and is based on the equations provided in Section 3. A more detailed explanation of how the clustering process is done and why it is useful is included in both Section 3 and the case studies descriptions.

o) From line 318 to 331, could you explain how to systematically conduct the procedures in Table? The interpretation of the procedures is not clear and not easily applied by new readers.

We concede that there is difficulty understanding the complexity of creating Figure 7 and its interpretation. We have added additional explanation of how the *DayFilter* process is executed in the simplified 2 day and 2 week examples in Section 3. The process outlined in that section is exactly the same as what was applied to the two case study datasets – albeit with many more daily profiles. We believe that the process and visualizations scale well with the amount of data and that is one the advantages of using the process. We enhanced Section 3 to the point where this should likely be clear to the reader.

p) For item 4.2, please do the same things as the first case study (item 4.1).

We have added additional context information and discussion for Case Study 2 to Section 4. However, for the sake of brevity, we intend not to go into as much detail with Case Study 2 as we did Case Study 1. The key intent of this second case study is to simply contrast the sankey diagram/heatmap between two datasets that differ in climate, location, use-type, and systems. Not as much investigation effort was placed in Case Study 2 due to this focus. We have clarified this intent so there would be no misunderstanding of the lack of content of the second case study as compared to the first.

q) In Fig. 10, what are 0, 1, and 2,3,4,5? Also, with the same paragraph, please give more details of three heating and cooling season since each case will match to the order of each number provided in Fig. 10. Please explain more clear for the second case study.

The numbers in the former Figure 10 (now Figure 12) refer to the automatically generate performance clusters extracted from the motif candidates using the k-means algorithm. These profiles were generated based on their raw-date feature similarity across each of the daily timestamps. We have added more content in Section 4.2 to discuss the interpretation of the results of each cluster.

r) For item 5 in discussion, I recommend you to combine it with item 4. Please discuss the results and what you found after you concluded the results for each case study in item 4.

We have gone into much more detail within the case studies regarding conclusions derived from the process. This includes the addition of more detailed discord analysis and comparison with outside influencing variables in Case Study 1. We believe these additions enhance the analysis of the case studies in a way that satisfies the reviewer's concern.

We have taken much of the content presented in the former Section 5 (Discussion) and put it in its own Section related to general Parameter selection analysis. We believe that this content is much more general as applied to the process as opposed to the case studies and thus, should have it own section.

A new Discussion Section was created as Section 6. It has been developed to account for much of the deficiency in detailed analysis of the process as it relates back to the Introduction. This section discusses the limitations of the process and the properly applied context in much more detail. We discuss the process as it relates to AFDD and the model calibration contexts introduced in the previous sections.

s) On line 417, how do you define the definition of data quality? It is based on the accuracy of measurement or fault-free data (without fault operations of machine affecting total power consumption). Also, what are the limitations of your application? For example, if data include fault operation data of building equipment, your application cannot identify or analyze what fault situations occur in building operations.

The definition of data quality has been further developed in the Discussion section. Data quality in our context refers to accuracy of the physical phenomenon occurring and the frequency and reliability of data collection storage and management. A pre-qualification of the case studies was done to ensure proper calibration of sensors and data management.

3 Review 3

3.1 General comments from reviewer

An interesting and useful application of a technique developed in another (perhaps more general) domain.

3.2 Response to general comments

Thank you very much comments – we're encouraged based on the response that this is an 'interesting and useful application'. Thanks for the comments on clarification of the SAX aggregation vs. clustering step. We agree that the intermediate steps needed to be clarified in more detail as the intent of the paper is that the reader would be able to implement the process on their datasets. Also, thanks for pointing out an obvious error in our analysis in interpreting the clustering for Case Study wrongly.

3.3 Revision items

t) It's not clear to me what the clustering step is for and how it differs from or compliments the SAX transformation and daily profile tagging/filtering. I got little sense that this step provided any added information either from the description in section 3 or the case studies in section 4. The discords and motifs in the case studies appear to provide information and have value but again, I don't see the value in the clusters. What does it mean to be in cluster 0 or cluster 1? Unless a case can be made that this step contributes to detecting abnormal or problematic operation; gives insight into building operation; or supports decisions, it should be left out.

Thanks for prompting us to provide more detail in the clustering step, which was lacking in explanation of purpose. We have added additional content to Section 3.4 to outline the motivation for the step – in brief, the purpose is to give the ability to further aggregate the motif candidates for the purposes of design-phase feedback after discords have been filtered out. For example, the SAX parameter inputs could be tuned to create many more motif candidates by simply increasing either A or W, which could produce 30+ motifs based on the diversity of patterns in the dataset. The clustering step provides a means of further aggregation to tune the number of patterns according to the level of user needs.

We have also further developed the motivation for using the process in the model calibration context which reinforces the usefulness of creating a certain number of performance profiles based on expert judgment.

u) While on clustering, for the first case study (Singapore building), I don't see how cluster 1 is "strongly prevalent on the weekends" (line 337); to my eyes it appears more prevalent on the weekdays. This should be clarified.

Thank you very much for pointing out this mistake! Cluster 1 was wrongly interpreted as being a majority 'weekend' day-type. It is holiday or half-day type of schedule for the school where only certain parts of the building are occupied such as when the teachers are working and the students are on holiday.

v) Section 5 could be expanded. The treatment of how the parameters impact results is welcome and necessary, but the discussion is mostly qualitative and does not assess how changes in these parameters and corresponding changes in results might change a user's interpretation of the data. This section comes across as a little unconvincing and its conclusions need greater support.

Section 5 was changed from being labeled as Discussion to its own dedicated section titled *Parameter selection analysis*. This new dedicated section goes into more detail of the 6 scenarios created to show the difference between the outputs when selecting reasonable ranges for the the SAX process parameters, SAX alphabet, A, and window size, W. The original version of this section simply illustrated the number of patterns created when selecting smaller windows or larger alphabets – which seemed to disadvantage the decision to select these scenarios as they would create more patterns than could be easily interpreted. The analysis is now offset with the presentation in the differences in variation within the patterns according to these two parameters. The trade-off is now more apparent in that smaller windows and larger alphabets gives more resolution to the process, thus creating more tightly grouped clusters and perhaps more effectively detecting less systematic discords.

In addition to illustrating the difference in the number of patterns generated, we also show the difference in the variance within each SAX word pattern based on the input parameter settings. This concept is shown in Figures 16 and 17 and accompanying text.

w) On line 405, use a different phrase than "significant statistical". This phrase implies that a rigorous and quantitative statistical test of significance was conducted - a standard and well defined technique - yet the only analysis was a visual inspection.

Thank you for this important clarification. We concede that this process is validated using many qualitative aspects of clustering validation. This section was meant to illustrate two common statistical metrics used for quantitative, internal clustering validation. Despite this effort, yes, we concede that *significant statistical* is not the best description of these experiments. Thus, we have removed this phrase and re-edited this section.

x) *The following should at least be discussed a full treatment need not be included in this paper:* Given the objectives of using the fewest parameters possible and "let the data itself speak to us" it would be helpful to have some sensitivity measure indicating how parameter changes cause interpretation or discord detection changes - with a low score on such a measure being more desirable. Some guidelines are offered and while these seem reasonable they are not rigorously supported - although I recognize that providing rigorous generalized rules may be difficult to do for this technique and application and thus would be a matter of experience in the field.

Thank you for this insightful suggestion. We believe that application of this process on multiple variables from larger datasets from many more buildings will provide the type of insight outlined in this comment. Each dataset the technique is applied to would need to have a ground truth match in order to test the accuracy of fault detection according to the discord analysis. Unfortunately, large datasets like those used in this study are rare – and even Case Study 2 in this paper is lacking in the information to perform an adequate efficacy study of detection. There are plans to continue the application of this process to many more buildings but this will take time.

If you are interested, we have begun to start a repository of raw datasets from commercial buildings that can be shared online. The dataset from Case Study 1 is a seed dataset for this study. If you're interested, please visit www.datadrivenbuilding.org or the poster paper recently presented "A seed dataset for a public, temporal data repository for energy informatics research on commercial building performance" at the Erasmus EnergyForum 2014, Rotterdam, Netherlands. June 19-20, 2014, which is Reference [60].

y) *The following are suggestions that might enhance the paper but not something I find absolutely necessary:* As motivation for this work, it might be helpful to discuss briefly how much performance gaps are due to faulty/discordant operation and how much is due to overly optimistic modeling/simulation during design, uncertainty, etc. Also interesting would be to quantify how much this approach narrowed a performance gap: are there any design-stage performance predictions for the case studies (although this may be more appropriate for a follow up study)?

This suggestion is fascinating for us as the process is meant to help with this distinction. However, no design phase models are available to us for these two case studies – a scenario that is very common. In fact, we have requested the energy model for Case Study 1 from the consultant who developed it 4 years ago for the construction phase. They were actually unwilling to share the model as they claimed it was their intellectual property! This is the despite the fact that the owner requested them to share it with us and had paid them a rather large fee to develop the model. Much more research in this area needs to be completed and we are focused on such topics. Unfortunately, this can't be accomplished for this study at this time.

z) Section 6, the conclusion could perhaps be more reflective of section 1.

The discussion section was completely rewritten and it now consolidates the perspectives gained in the study as compared to the literature review and introduction in a much more detailed way. We discuss the process as it relates to AFDD and model calibration.