Leakage Detection with Non-Negative Decomposition Models Constrained with Single Fixed Component

Sverrir Heiðar Davíðsson
DTU Compute (Department of Applied Mathematics and Computer Science)
Danmarks Tekniske Universitet
Kgs. Lyngby, Denmark
SverrirHD@gmail.com

ABSTRACT

This paper proposes an alteration to two non-negative decomposition models, non-negative matrix factorization NMF and non-negative parallel factor analysis (PARAFAC), to estimate leakages in water distribution systems. The proposed alteration is a change to the alternating least squares algorithms used to fit both models such that a single component (or factor) remains fixed and flat. This pattern is supposed to resemble the expected pattern of leakages. The paper describes three experiments that show: 1) a flat weekly pattern is the latent representation of leakages 2) The alteration enables both models to extract the flat pattern where the unaltered methods were previously unable to do so and the effect is just as accurate of an estimation of leakages 3) The method generalizes well to dissimilar usage patterns and is able to detect multiple known leakage repairs as a proxy for leakage detection.

CCS CONCEPTS

• **Unsupervised learning**~*Source separation* • **Factorization methods**~*Non-negative matrix factorization*

KEYWORDS

Non-negative matrix factorization, PARAFAC, leak detection, pattern analysis, unsupervised, water distribution systems.

1 INTRODUCTION

One of the biggest challenges utility companies face all over the world is the large discrepancy between water put into water distribution systems (WDS) and the water billed to consumers. This lost water is sometimes called non-revenue water (NRW) and about 80% of NRW is estimated to be caused by leakages from pipes. The total cost of NRW worldwide is conservatively estimated to be about $14 billion per year. In developing countries NRW is estimated at 35% of total system input and in developed countries it's estimated at 15%. [4]

The three main challenges in reducing total NRW in WDSs, before the leakages can be repaired, are: leak detection, leak estimation and leak localization. These are all very related and hard to separate from each other and none can be considered more important than another. Size estimation is especially important for to prioritize repairs, since often more leakages are known than can be dealt with at once and leakage repair can be very resource intensive and costly.

Leak localization methods can be categorized broadly into two categories, external and internal systems. External systems use sensors attached to the pipes, such as acoustic and vibration sensor, to trigger alarms when certain criteria are met. Internal systems use analytical techniques to process and analyze data from external hardware. Most of these methods focus on leak detection rather than estimation. [2]

In this paper we consider an approach which touches on all three of the main challenges in reducing total NRW and would be categorized as an internal system method. The method can only locate the leak to the area which is bounded with the sensor(s). The main contribution of this paper is introducing a form of leakage estimation based on a principle of semi-blind source separation using methods that have been modified specifically with leak detection in water utilities in mind. The method is 'semi-blind' since we introduce into the models prior knowledge specific to WDSs, such as non-negativity, a single known pattern of behavior and the period size of the patterns. The approach is designed for flow measurements for a bounded region where all input is measured.

The two models considered in the paper each constrain the patterns in a slightly different way. The paper describes an alteration to the algorithm which both models use to fit the data. The alteration constrains the models to estimate new patterns subject to one of the patterns, the leakage pattern, being flat*. An increase to the leakage patter corresponds to an equal increase in measured flow for all hours of the day/week. This is in contrast to increases due to public being unequally distributed over the week, since people's usage is bound to the hours they are awake, are at home and their behavioral patterns. In order to discern this type of increase from other types of increases, we must first have an estimate of the other patterns. For this purpose we chose unsupervised models, since they will be able to find patterns even if they are unique to each region.

2 METHOD

2.1 Models

The two models we consider are: non-negative matrix factorization (NMF) and parallel factor analysis (PARAFAC).

*For further discussion about the flat pattern and limitations, see appendix section 5.1

PARAFAC will be used with a non-negativity constraint and referred to as non-negative PARAFAC. Non-negative PARAFAC may be viewed as an extension of NMF from matrices to tensors. The effect of this difference is the additional constraints applied to the patterns. NMF only constrains the patterns to be non-negative and weekly in length while PARAFAC constrains the patterns to be daily, but allows for variation in the scale of the daily patterns for each day of the week, so a weekly pattern from PARAFAC is a series of differently scaled daily patterns of the same shape.

2.1.1 NMF

The NMF implementation is based on the scikit-learn NMF implementation using the Frobenius loss as a cost function.[6] The objective of NMF with this loss is to solve

$$(\overline{\boldsymbol{W}}, \overline{\boldsymbol{H}}) = argmin\ \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \quad (1)$$

where $\boldsymbol{X} = [x_{ij}]$ is a non-negative $m \times n$ matrix $\boldsymbol{W}$ is a non-negative $m \times r$ matrix and $\boldsymbol{H}$ is a non-negative $r \times n$ matrix and r is the number of non-negative components.

2.1.2 Non-negative PARAFAC

The non-negative PARAFAC implementation is based on the Tensorly implementation. [7] This method also uses the Frobenius loss as a cost function and the objective is to solve

$$(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \overline{\boldsymbol{C}}) = argmin\|\boldsymbol{X} - \textstyle\sum_{i=1}^{r} \boldsymbol{a_i} \otimes \boldsymbol{b_i} \otimes \boldsymbol{c_i}\ \|_F^2 \quad (2)$$

where $\boldsymbol{X} = [x_{ijk}]$ is a non-negative $l \times m \times n$ tensor $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ are non-negative matrices with elements a_{ir}, b_{jr} *and* c_{kr} where r is the number of non-negative components and $\otimes$ is the tensor product. The resulting output from the tensor products has the same shape as the original tensor, $\boldsymbol{X}$.

2.2 Patterns

We refer to the $\boldsymbol{H}$ matrix in (1) and the $\boldsymbol{B}$ and $\boldsymbol{C}$ matrices in (2) as the *pattern matrices* and the $\boldsymbol{W}$ matrix in (1) and $\boldsymbol{A}$ matrix in (2) as the *abundance matrices*. We call the rows of H and every pair of b_i and c_i vectors *patterns*. We call to the rows of $\overline{\boldsymbol{W}}$ in (1) and $\overline{\boldsymbol{A}}$ in (2) the *estimated abundance* (of the patterns).

When we describe patterns in this paper we are referring to, in the case of NMF, the rows of H, and, for PARAFAC, every pair of corresponding vectors b_i and c_i. In the case of NMF, we look for vectors/patterns such that a linear combination of them approximates every week with minimal error. For PARAFAC, each week in the tensor $\boldsymbol{X}$ is stored as 7×24 matrix and so the outer products of a pair of b_i and c_i vectors creates a pattern in the shape of a 7×24 matrix. These matrix patterns are then scaled with the corresponding weights from a_i to minimize the error of reconstruction.

2.3 Solvers

Both implementations are based on the alternating least squares algorithm, which is described in (ALGORITHM 1), to solve the objective function. The NMF implementation uses HALS (hierarchical alternating least squares)[3] and PARAFAC uses an alternating least squares algorithm [1].

ALGORITHM 1: Alternating Least Squares

```
Initialize matrices with random non-negative values
selected matrix is abundance matrix
while convergence criteria is false, do
        minimize objective function wrt. selected matrix
        selected matrix is next matrix
end
```

2.3.1 Proposed change to solvers

The proposed change assumes leakages have a flat pattern. (*For further discussion on cases where this assumption may not apply and what other possible causes of continuous flow may exist in the system, see appendix, section 5.1.*)The proposed change is described in (ALGORITHM 2).

ALGORITHM 2: Modified Alternating Least Squares

```
Initialize matrices with random non-negative values
Flatten first component of pattern matrices
selected matrix is abundance matrix
while convergence criteria is false, do
        minimize objective function wrt. selected matrix
        Flatten first component of pattern matrices
        selected matrix is next matrix
end
```

Flattening in (ALGORITHM 2) refers to replacing all weights of the vector with the same value. This value may be chosen arbitrarily since it may be considered a free parameter in the reconstruction as long as the scale of the corresponding pattern abundance is not fixed. We chose to scale the flattened algorithm such that its mean value was the same as the mean value of the other two patterns. This was done for ease of comparison, since the.

2.3 The data

In this paper, we use two time series, which both contain hourly means of flow measurements. One is a residential neighborhood called Skerjafjörður, the other is a stable area called Víðidalur. These were selected for their dissimilarity in nature of usage and also since in both cases: all incoming water goes through the measured pipe, both time series had apparent leakages and repairs and there were no large gaps in data.

2.4 Experiments

For NMF, the data is reshaped into a $(w\ x\ h_w)$ matrix, where w is the number of weeks (samples) in the data and h_w is the number of hours (variables) in a week. For PARAFAC the data is reshaped into a tensor with the shape $(w \times\ d_w \times h_d)$, where w is the number of weeks in the data and d_w is the number of days in a week and h_d is the number of hours in a day.

For unsupervised methods such as these, the number of components (*or factors*) is an important parameter. This can be selected in several ways but there is no obvious way to choose them that would suit the objective of these experiments. For simplicity and ease of interpretation we will consider three components for all the experiments.

2.4.1 Experiment 1 – Show the leakage pattern is latent

The first experiment is to show that a flat pattern is latent in the data and that changes in leakages due to repairs correspond to changes in its abundance estimate. If this can be shown using the unaltered methods, we have more reason to believe that our understanding of leakages is correct and that the modified methods will be able to estimate leakages in more neighborhoods than the unmodified method.

Both methods will be used on both datasets. The training will be done using all available data and then we will consider the abundance estimates, specifically a few weeks before and after the largest known repair in either dataset.

2.4.2 Experiment 2– Reproduce results with modified methods

Since the results of experiment 1 show, as seen in section 3.1, the abundance estimates for the flat pattern changed as expected around the same time as the large repair took place. Now it's time to show that the modified method is capable of producing the same results for the same dataset as well as the other dataset, which the unmodified method was not able to do.

This should demonstrate the modified methods ability to generalize this leakage-estimating ability to neighborhoods with dissimilar patterns.

2.4.3 Experiment 3 – Estimate repair detection capability

Now we have seen that both modified methods are able to produce good estimates of leakage abundance in either system around the same time as the largest repair took place in either dataset. What we want to show now is that these two large leakage repairs were not the only case where the methods can achieve these results.

In this final experiment we intend to estimate the abundance of the leakage pattern for a much larger period of time. When leakages are repaired, we should see a sudden decrease in the mean value of the leakage estimate. This change is called a step and there are method used to detect these changes which together are called step detection. [5] The specific method we will use involves convolving the signal with a step function as a filter, which we choose to be 10 weeks in size. This filter has values of 1 for the first 5*168 weight and values of -1 for the last 5*168 weights. This is similar to how the gaussian derivative is used for edge detection in image analysis. If there is no change to the signal during this window, then the convolved signal will be close to 0. If there is an abrupt change in the mean, we will se a global minimum if the mean drops, as with repairs, and a global maximum if the mean suddenly increases, as with sudden leakages. [5]

We will observe how well the method was able to detect the times a repair has been made. Note that we will be looking at a 40 day window centered at the date of the report, since this date often differs from the date of the repair by up to several weeks. (*See appendix section 5.2 for further elaboration of the issues with these dates*). There were much fewer and less reliable reports in the stable area, so we focus only on the residential neighborhood.

3 RESULTS AND DISCUSSION

3.1 Experiment 1

Only non-negative PARAFAC produced the flat pattern on the residential dataset. Neither did so on the other dataset. This however is enough to show that a flat pattern is naturally latent in the data and that changes in its abundance estimate follow changes in the amount of leakages in the system (i.e. repairs). (See appendix section 6.5 for results from all neighborhoods).

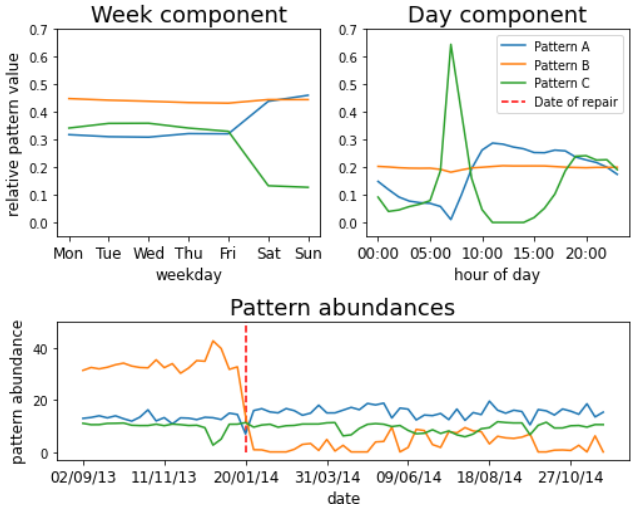

Figure 1: Result from unchanged non-negative PARAFAC on residential neighborhood data

3.2 Experiment 2

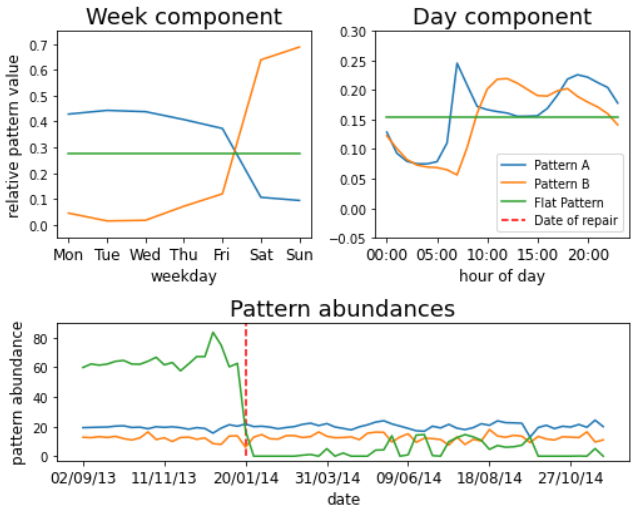

Figure 2: Result from non-negative PARAFAC with proposed change on residential neighborhood data

Comparing the results week and day components of the patterns in Figure 2 with those in Figure 1, we can clearly see that the shape of the non-flat patterns did not change much despite the changes made to the method.

Both modified methods were both able to mimic the desirable behavior that emerged in experiment 1. This indicates that the modified methods will be able to estimate leakages in many more neighborhoods than the previous unmodified methods

Looking at Figure 3 we see that the estimated abundances of the flat pattern were almost identical for the two changed methods. Because of this, we will focus only on NMF for the next experiment.

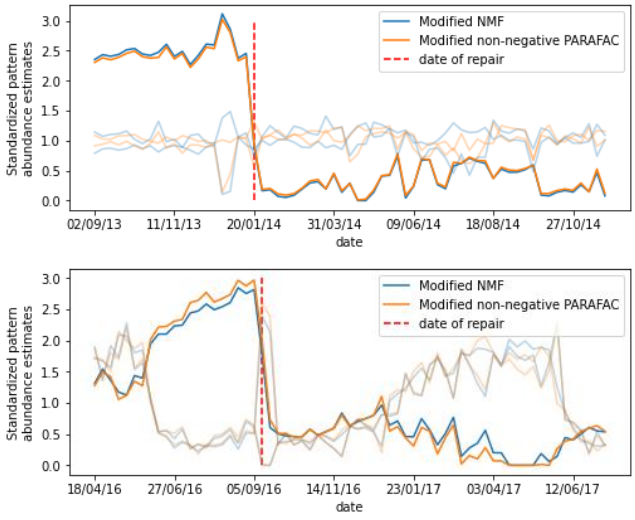

Figure 3: Leakage abundance estimates using modified NMF and modified non-negative PARAFAC. A repair in the Skerjafjörður residential neighborhood (top) and a repair in the Víðidalur stable area (Bottom). The darker lines indicate the leakage pattern estimates and the lighter lines indicate the non-leakage pattern estimates.

3.3 Experiment 3

Figure 4 shows that that almost every report in the residential neighborhood is within 20 days from a local minimum of the convolved signal. This is even stronger evidence that the method is working as intended.

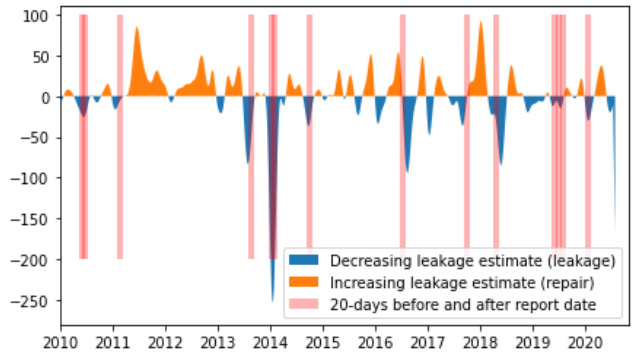

Figure 4: Convolution of leakage estimate with a windowed step function. The translucent red bars indicate 20 days before and after a report which is usually accompanied by a leakage repair.

4 CONCLUSION

In this paper we proposed the use of NMF and non-negative PARAFAC for leakage estimation by indirectly constraining the models to include a flat factor by modifying the algorithm used to fit the models. We showed first that non-negative PARAFAC was able to extract the flat leakage pattern in a completely unsupervised manner for one of the datasets. We also showed that this pattern was a good latent representation for leakages by showing that its abundance estimate matched leakage repairs.

The most important results were showing that both of the altered methods were able utilize the flat pattern to estimate the amount of leakages in both datasets. This result indicates that the altered method can be used on much more diverse data than the unaltered unsupervised methods.

Further development of the method could be done to incorporate evolution of patterns over time, like dynamic filtering in digital signal processing. It is likely that better results could be achieved with better parameter tuning using number of components and regularization parameters. The method as it stands is only though to work for systems where the total input is measured.

REFERENCES

[1] de Almeida, André & Favier, Gerard & Carvalho Lustosa da Costa, Joao Paulo & Mota, joão. 2016. Overview of Tensor Decompositions with Applications to Communications. *SIAM Review*, (Aug. 2009), 455–500. DOI:https://doi.org/10.1137/07070111X

[2] Awang Lah, Airull Azizi & Dziyauddin, Rudzidatul & Md. Yusoff, Nelidya. 2018. Localization Techniques for Water Pipeline Leakages: A Review. *International Journal of Integrated Engineering. 10.* https://doi.org/10.30880/ijie.2018.10.07.028

[3] Cichocki, Andrzej and Phan, Anh-Huy. 2009. Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. IEICE Transactions. 92-A. 708-721. DOI:http://doi.org/10.1587/transfun.E92.A.708

[4] Kingdom, Bill; Liemberger, Roland; Marin, Philippe.2006. The challenge of reducing non-revenue water (NRW) in developing countries - how the private sector can help : a look at performance-based service contracting (English). *Water Supply and Sanitation Sector Board discussion paper series no. 8;* Retrieved November 30, 2020 from http://documents.worldbank.org/curated/en/385761468330326484/The-challenge-of-reducing-non-revenue-water-NRW-in-developing-countries-how-the-private-sector-can-help-a-look-at-performance-based-service-contracting

[5] Picard, D. 1985 Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4), (December 1985), 841-867. DOI: https://doi.org/10.2307/1427090

[6] Scikit-learn, 2020. sklearn.decomposition.NMF — scikit-learn 0.23.2 documentation. Retrieved November 30, 2020 from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

[7] Tensorly. 2020. tensorly.decomposition.non_negative_parafac. Retrieved November 30, 2020 from http://tensorly.org/stable/modules/generated/tensorly.decomposition.non_negative_parafac.html#tensorly.decomposition.non_negative_parafac

5 Appendix

5.1 Further analysis of leakage pattern

5.1.1 Other causes of continuous flow in the system

In the paper, the pattern of leakages is stated to be flat but, for sake of being thorough, alternative causes of the same pattern were considered. There exist multiple possibilities of these non-leaks which exhibit the behavior of continuous flow. This usage would be indistinguishable from leaks in the system and therefore a rough investigation was made into possible causes of such usage.

A former specialist in leakage localization at the utility company Veitur ohf. noted that the most likely cause of continuous flow in residential neighborhoods was toilets with a broken flow regulator. This was considered a likely culprit after a survey was conducted some years ago to see how many people had continuous flow in their toilet bowls. The result was an estimated 1 out of every 20 toilets which had continuously running water. There was also an estimated 1 toilet per 4 people in a residential area. This along with an experiment conducted within the company to see how much water flowed from continuously running toilets, which was about 0.03 liters per second lead to a rough estimate of 0.000625 [l/s] per person in a given neighborhood. In a ~800-person neighborhood like Skerjafjörður, this would amount to about 0.5 [l/s].

The specialist also noted other possible causes such as garden tools and residential structures such as swimming pools, water fountains, garden streams, etc. According to regulation these things should not be continuously flowing water but should instead have some sort of recycling system or regulation to prevent it. The specialist noted that in his experience people were often quick to ignore regulation in favor of conveniency, and especially so since cold/drinking water is not priced from meters for residential users, as is the case with hot water is in Iceland, but rather estimates of usage using the size of a house's footprint and the measured mean usage of the neighborhood or similar neighborhoods in the case of non-metered neighborhoods.

5.1.2 Non-flat nature of real leakages and problems with true leakage estimation.

In the paper we use a completely flat pattern to represent leakages, but leakages are subject to changing characteristics of their environment as well as changes in the internal pressure of the pipe. In systems with old and/or narrow pipes the pressure in the system can drop quickly with increased demand. This causes a drop in the flow of water out of leakages because the flow from leakages is a function of pressure, size and shape of opening.

The most obvious reason why the pattern of a leakages is not necessarily flat is because leakages very often grow over time. This can be caused by erosion of the pipe, the surrounding soil or simply increases in pressure.

Another indication that the leakage pattern isn't necessarily flat can be seen in the results of the unmodified non-negative PARAFAC profiles. The almost flat profile had a very slight but significant drop at the same time as peak demand in that neighborhood. It could also have been coincidental, but since the phenomenon is known, it's less likely so.

Changes to soil characteristics due to pressure from objects on the surface, water saturation or other reasons might also cause the leakages to have a non-constant flow of water over time.

5.2 Problems with report data

A slight problem with the approach used in experiment 2 was that the dates used as 'repair dates' are made using imperfect data. The dates come from two types of reports: one type is made after leakages localization attempts have been made in an area. Following these, repairs will sometimes be made, and the time of these repairs is sometimes days to several weeks after the report has been made, depending on the priority of the repair. The other type of report is an operations report, documenting the repair operations performed on a given date. These will often be more accurate, but the date the report is made is sometimes a few days or up to a few weeks after the operations themselves.

To take this into consideration, we visualize the repairs as a 40-day window centered on the date given and see if we have an estimated drop in the leakage estimate during the same period. Another problem is that sometimes there will be a report made about a leakage search which found a leakage but there is no follow up about whether a repair was eventually made.

5.3 Experiment 1 complete results.

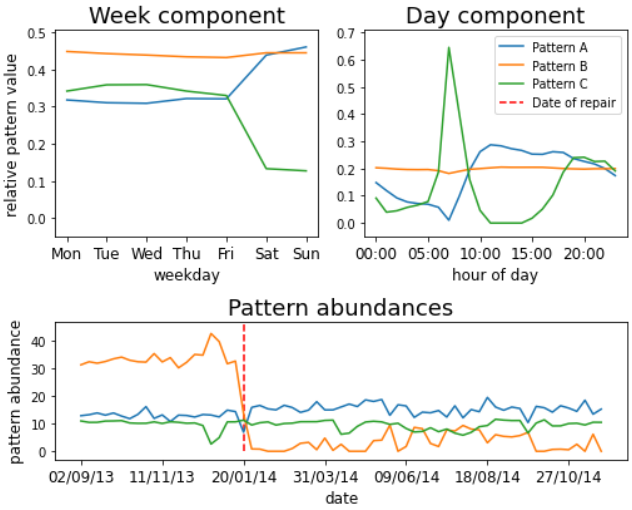

Figure 5: non-negative PARAFAC results from residential neighborhood

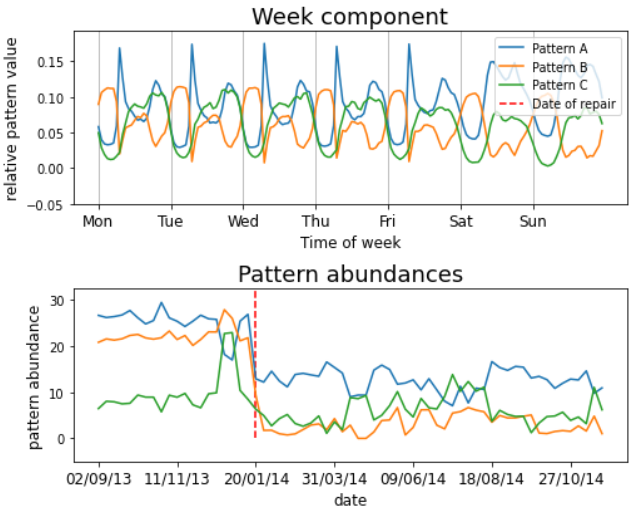

Figure 6: NMF results from residential neighborhood

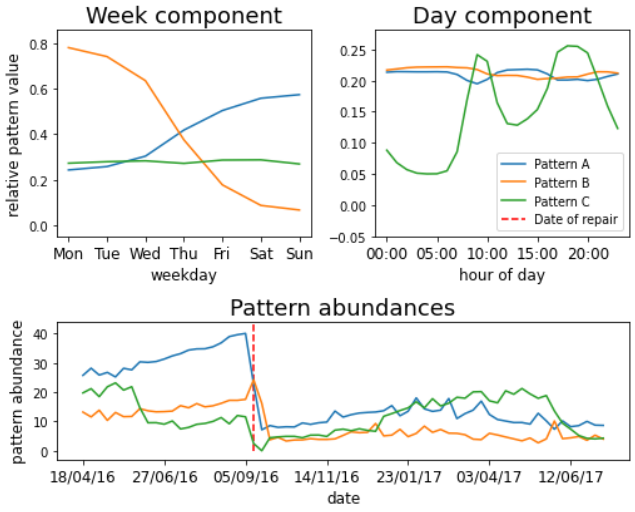

Figure 7: non-negative PARAFAC results from stable area

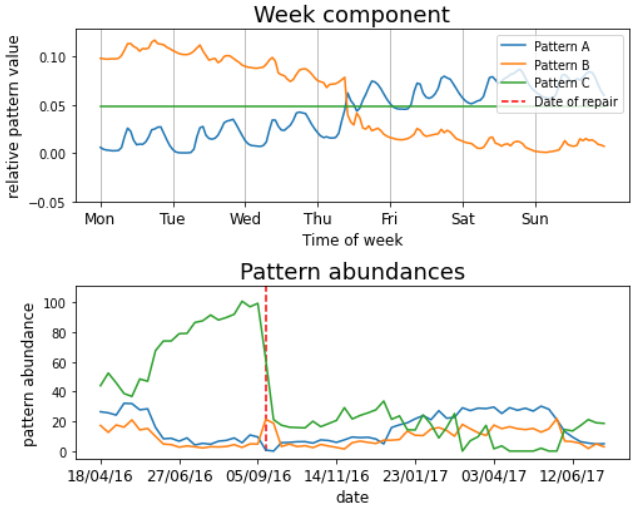

Figure 8: NMF results from stable area

5.4 Experiment 2 complete results

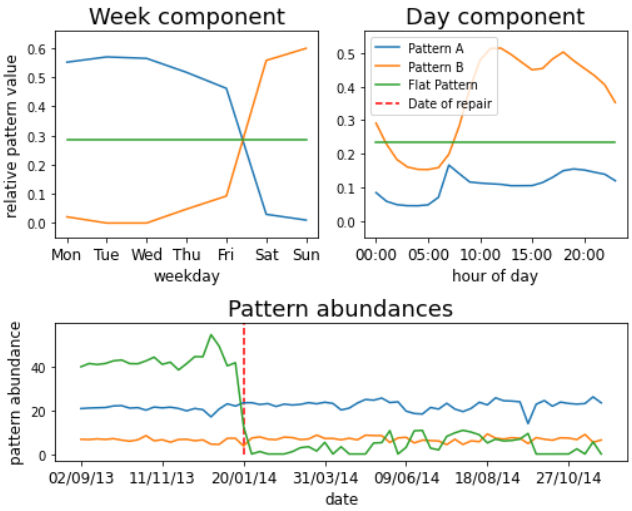

Figure 9: Result from non-negative PARAFAC with proposed change on residential neighborhood data

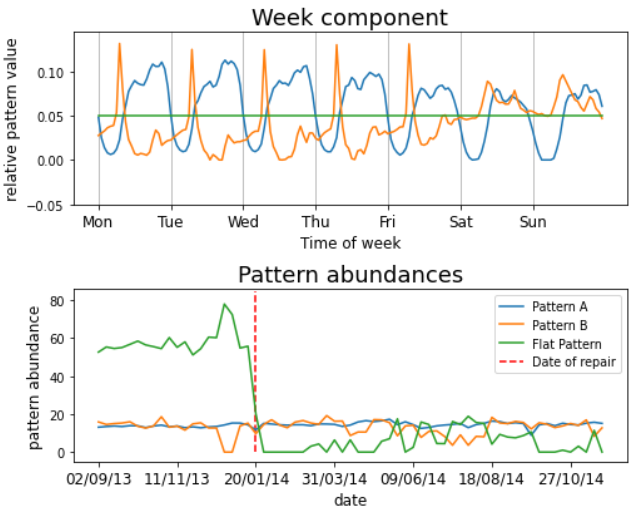

Figure 10: Result from NMF with proposed change on residential neighborhood data

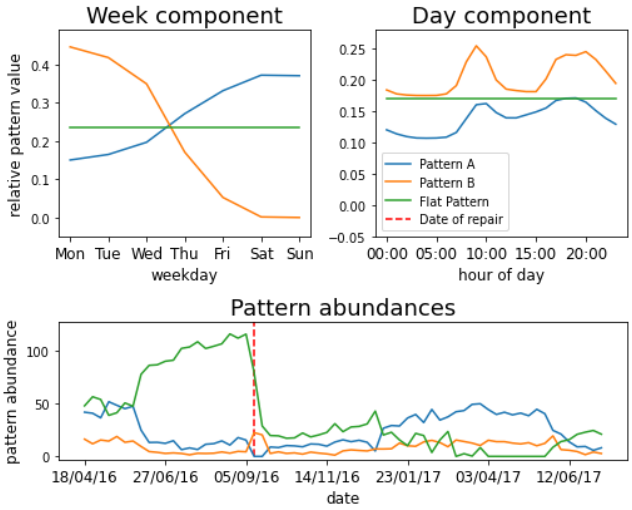

Figure 11: Result from non-negative PARAFAC with proposed change on stable area data

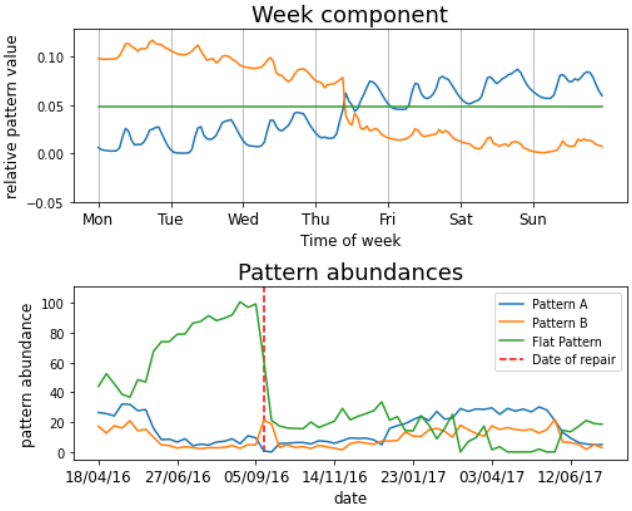

Figure 12: Result from NMF with proposed change on stable area data.