

Using Dynamic Time Warping Self-Organizing Maps to Characterize Diurnal Patterns in Environmental Exposures

Kenan Li (✉ kenanl@usc.edu)

University of Southern California <https://orcid.org/0000-0002-4641-6699>

Katherine Sward

University of Utah

Huiyu Deng

University of Southern California

John Morrison

University of Southern California Keck School of Medicine

Rima Habre

University of Southern California Keck School of Medicine

Meredith Franklin

University of Southern California Keck School of Medicine

Yao-Yi Chiang

University of Southern California

Jose-Luis Ambite

University of Southern California

John P. Wilson

University of Southern California

Sandrah P. Eckel

University of Southern California Keck School of Medicine

Methodology

Keywords: Dynamic Time Warping, Self-Organizing Map, time series clustering, pattern discovery, asthma

Posted Date: October 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-87487/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2021. See the published version at <https://doi.org/10.1038/s41598-021-03515-1>.

Abstract

Background

Advances in measurement technology are producing increasingly time-resolved environmental exposure data. We aim to gain new insights into exposures and their potential health impacts by moving beyond simple summary statistics (e.g., means, maximums) to characterize more detailed features of high-frequency time-series data.

Methods

This study proposes a novel variant of the Self-Organizing Map (SOM) algorithm called Dynamic Time Warping Self-Organizing Map (DTW-SOM) for unsupervised pattern discovery in time series. This algorithm uses DTW, a similarity measure for sequential data that optimally aligns interior patterns, both as the similarity measure and for training the neural network.

Results

We applied DTW-SOM to a panel study monitoring indoor and outdoor residential environmental exposures for 10 patients with asthma from 7 households near Salt Lake City, Utah; each patient was followed for up to 373 days. Compared to other SOM algorithms using Euclidean distance, the DTW-SOM algorithm maintained the topological properties of the input time series and generated more detailed diurnal patterns. We observed seasonal patterns in outdoor temperature and distinct patterns of indoor peak PM_{2.5} exposure, which was likely linked to both combustion sources and days with increased inhaler usage.

Conclusions

The new algorithm, DTW-SOM, better preserved the topology relationship of time-series data and better summarized time-series patterns as compared to the original version of SOM.

1. Background

A standard approach in air pollution health impact studies is to relate continuously varying ambient exposures to health outcomes using exposure history summaries such as 24-hour averages (Auchincloss et al., 2008; Liu, Krewski, Shi, Chen, & Burnett, 2003). Current daily air quality regulations in the United States (US) are based on the Environmental Protection Agency's (EPA) Federal Reference Method (FRM), which collects 24-hour integrated samples and the Federal Equivalence Method (FEM), which collects hourly samples at Air Quality System (AQS) network monitoring sites. The EPA releases hourly, daily, and annual data, where the daily and annual summaries are sometimes averages of shorter-term measurements. By using averages we may miss key short-term temporal variability in exposure that affects health differently than long-term averages. For example, (Delfino, Zeiger, Seltzer, Street, & McLaren, 2002) examined the impacts of air pollution "peaks" in a study of children with asthma and

found stronger evidence for an association of asthma symptoms with the daily 1 hour maximum of outdoor PM_{10} than the 24-hr PM_{10} averages. The development of personal and stationary air pollution monitors and low cost air sensors provides highly time-resolved exposure data and new approaches are needed to summarize these data and explore potential health impacts.

Here, we focus on a new method to identify typical diurnal patterns in exposure time-series data, especially under circumstances with slight time warping or shifting. For example, weekday NO and NO_2 have a typical diurnal pattern with two periods of elevated levels, with the first related to the morning traffic peak and the second related to the evening traffic peak. The morning peak of NO_2 normally appears 1–2 hours after the NO peak (Han et al., 2011). However, due to day-to-day variation in meteorology or traffic patterns (e.g., accidents), the weekday traffic peaks of a given pollutant (and lag between pollutant peaks) may not always occur at precisely the same time. We propose a new method that can identify typical diurnal patterns in exposure time series under temporal non-stationarity. Our method accounts for variations in the timing of the diurnal patterns by comparing segments from different time periods when comparing two daily time series.

To discover typical patterns in time-series data, two broad approaches can be taken: (1) *supervised* time-series classification when *a priori* grouping information (e.g., health response) is available or (2) *unsupervised* time-series clustering. Supervised time-series classification targets patterns of exposure, for example, to discriminate between days where study participants did or did not have asthma exacerbation. Unsupervised time-series clustering characterizes observed exposure patterns independently of their association with health outcomes and can be used to address topics such as the diurnal patterns observed in daily indoor or ambient pollution exposures or how frequently these patterns occur. Patterns identified in an unsupervised clustering analysis can be later included as exposure measures in health models.

Time-series clustering has been well studied over the past two decades, as summarized in a recent review by (Aghabozorgi, Seyed Shirkhorshidi, & Ying Wah, 2015). The approaches used for time-series clustering incorporate methods for assessing the similarity between time-series. Similarity metrics can be conceptualized as mathematical expressions that indicate the cost of transforming one time-series into another or the inverse of the distance between two time-series (Cleasby et al., 2019). Simple Euclidean distance is one of the most widely applied similarity metrics. However, by definition, its elementwise alignment makes it unable to capture the similarity of shapes with small distortions in the time axis. Diurnal patterns in air pollution—which may impact human health—arise from complex processes, so it is important to allow for small distortions over time (e.g., due to day-to-day variation in meteorology). Dynamic Time Warping (DTW) is a method that allows for elastic shifting of the time axis to detect similar shapes with different phases (Berndt & Clifford, 1994). Many temporal proximity-based clustering methods use DTW as a similarity measurement (Kremer, Gunnemann, & Seidl, 2010; Niennattrakul & Ratanamahatana, 2007; Warren Liao, 2005).

A Self-Organizing Map (SOM) is a clustering algorithm (Ritter & Kohonen, 1989) frequently applied in the exploratory phase of data mining. SOM transforms the input space onto a lower-dimensional (typically two-dimensional) gridded space effective for visualizing and exploring the properties of the input data. The original SOM algorithm uses Euclidean distance as a similarity metric. Since Euclidean distance is ill-suited to characterize misaligned sequential data, several previous studies have refined the SOM algorithm by replacing it with DTW (Juhasz, 2009; Juhász, 2007; Okada & Hasegawa, 2008; Scepi & Romano, 2006). However, these previous studies only replaced Euclidean distance with DTW in the matching phase of the algorithm but retained Euclidean element-wise alignment in the training phase of the algorithm which produces weights representative of a typical diurnal pattern. As a result, these previously proposed modifications to SOM are incomplete in their treatment of similar but misaligned patterns in sequential data and may produce suboptimal results.

In this paper, we propose a new Dynamic-Time-Warping Self-Organizing Map (DTW-SOM) algorithm that uses DTW as the similarity measure in both the matching and training phases of SOM. Thus DTW-SOM has the ability to better match similar patterns in time series with temporal misalignment as well as the potential to better characterize typical diurnal patterns. This novel methodological work was inspired by an application in environmental epidemiology, and we apply DTW-SOM to identify diurnal patterns in the residential particulate matter, temperature, and relative humidity exposures of patients with asthma.

2. Methods

2.1 Dynamic Time Warping

DTW detects and matches the internal patterns of two time series of the same size by calculating a two-dimensional distance matrix with all possible pairwise Euclidean distances between time points (Figure 1). DTW alignment is determined by finding the shortest path (the red line in Figure 1) that minimizes the overall combined values of the distance matrix under: (1) a boundary condition in that the path starts from the top-left and ends at the bottom-right corner to ensure that the alignment does not only partially cover subsequences; (2) a monotonicity restriction which requires that the path cannot go back in time to ensure that the internal patterns will not be repeatedly used in alignment; and (3) a continuity restriction that does not allow the path to break in time to ensure that no internal patterns are omitted. The DTW distance is then calculated by summing the Euclidean distance values along the shortest path. DTW can still be applied in cases with missing data because the matrix does not have to be square.

DTW gives a nonlinear (elastic) alignment between two time series. As an example, Figure 2 shows two synthetic signals with different baselines and a temporally misaligned (but otherwise identical) internal sine curve pattern. In subplot A, Euclidean distance assigns a vertical and element-wise alignment of all timestamps. Thus, it misses the similar—but temporally misaligned—patterns, although it does capture the relevant information on the different baselines. The difference between the two synthetic signals is 3, and the Euclidean distance is 3.05, which is very close to the true difference at baseline (time 0). In subplot B, DTW alignment elastically matches most parts of the sine curves, and the DTW distance is

0.39, which is much lower than the distance at baseline. The low DTW distance reflects the fact that the two synthetic signals share an identical pattern, under proper alignment.

DTW is computationally intensive with a quadratic time and space complexity, $O(N^2)$, that limits its usage on large time-series. However, many optimization techniques such as lower bounding, early abandoning, run-length encoding, bounded approximation and hardware optimization were proved to be useful in running DTW more efficiently and effectively (Mueen, Keogh, & Assoc Comp, 2016). Fast-DTW (Salvador & Chan, 2007) is an approximation of DTW that has a linear time and space complexity, $O(N)$. Fast-DTW relies on three key operations. First it “coarsens” the data into smaller time series with coarsened time resolution. Second, it finds the minimum-distance warp path at the coarser resolution. Third, it refines the warp path through local adjustments at finer resolutions. Although a previous research claimed that Fast-DTW is generally slower than the exact DTW in many realistic data mining application (Wu & Keogh, 2020), we found it took ~70 seconds for training SOM once with one data sample (i.e. 24 hours of data at a minute-level resolution with $60 \times 24 = 1,440$ timestamps) using standard DTW and 1 second for using Fast-DTW on a standard laptop computer (Intel Core i9-8950HK CPU @ 2.90 GHZ, 64.0 GB RAM, x64-based processor) in test runs on our data. We also found DTW and Fast-DTW results to be only <1% different in terms of distances and have the same warping path in most of the test runs. So, we applied Fast-DTW as a computationally efficient approximation to DTW in our DTW-SOM algorithm.

2.2 Self-Organizing Map

SOM is a type of unsupervised neural network with only two layers: one input layer and one output or mapping layer. The neurons in the input layer and the mapping layer are fully connected, which means each neuron in the input layer is connected to each of the mapping neurons and vice versa. However, the weight concepts in SOM have different meanings than in standard neural networks. Each weight on a connection in SOM represents the similarity between the connected mapping neuron and the input neuron. At each iteration of training, SOM searches for the mapping neuron whose weights are most like the input data (input vectors). The neuron with the best result is called the Best Matching Unit (BMU). This training regime is called competitive learning as opposed to error-correction learning in other standard neural networks. Then, the weights on the connection between the input vector and its BMU will be adjusted to make them closer. The weights of neighbors of the BMU are also modified closer to this input vector. The search radius for neighbors of BMU decreases at each training iteration. The training process stops either at the end of a pre-defined number of iterations or when the trained weights converge to a predefined threshold. At the end of the training, each neuron in the mapping layer represents a cluster center, and the observations that belong to that neuron/cluster are the ones having the shortest distances to the weights connected to that neuron.

SOM has been widely applied in time-series clustering (Cherif, Cardot, & Bone, 2011). In the field of air pollution epidemiology, SOM has been used to classify daily levels of several particulate and gaseous air pollutants into a set of multipollutant profiles, effectively identifying groups of days with similar pollutant mixture patterns (Pearce et al., 2014). This particular study sought to identify patterns in pollutant

mixture co-occurrence, so a data sample in their study consisted of daily average values of 10 ambient air pollutants from a US EPA Air Quality System (AQS) monitoring station in Atlanta over 8 years from 2000 to 2007. In this study, we sought to use SOM to discover the typical *diurnal* patterns for a single exposure variable (e.g., PM_{2.5}), such that our data sample consists of PM_{2.5} concentrations at each minute of a day from X sites/homes and Y days. Figure 3 shows the conceptual differences in the application of SOM between the two studies.

2.3 Dynamic Time Warping Self-Organizing Maps

Our proposed DTW-SOM has two key differences from the original SOM. First, inspired by previous studies (Juhasz, 2009; Juhász, 2007; Okada & Hasegawa, 2008; Scepi & Romano, 2006), we replaced the Euclidean distance similarity measure with DTW distance so that: (a) the BMU of a data sample is defined as the neuron in the output space with the minimal DTW distance to the data sample; (b) the distance-decay kernel function uses DTW distance; and (c) the weights of the neighborhood neurons of the BMU are updated under the kernel function using DTW distance. Second and most importantly, we developed a novel training regime for DTW-SOM. In the original SOM, for a neuron r , the rule for updating weights W_r corresponding to input x is given by the following equation:

$$W_r^{new} = W_r^{old} + \varepsilon \cdot h_{rs} \cdot (x - W_r^{old}) \quad (1)$$

where ε is the learning rate, h_{rs} is a distance-decay kernel function between neuron r and BMU s , and as the distance between r and s declines, h_{rs} increases. When r equals s , h_{rs} is at its largest. From Equation 1, we can see that the new weights are weighted means of the old weights and the input x under Euclidean alignment. However, if the BMU s is determined by DTW distance, this updating rule will adjust the weights of r towards an input x using Euclidean alignment, which is very likely not the one with the minimal Euclidean distance. In order to solve this issue, for DTW-SOM we revised Equation 1 to update weights W_r using DTW alignment. One of the major challenges in the revision is that the new weight of each timestamp t cannot always be directly calculated (Figure 4). The DTW-SOM updating rule is shown by the pseudo code in Table 1.

Table 1 Pseudo code of DTW-SOM updating rules

Calculate the DTW distance between input x and neuron r .

Let **best_path** denote the DTW alignment that pairs the timestamps in x with the timestamps in r .

Let **x_cords** and **y_cords** denote the coordinates of the weighted mid-points of each DTW pair in **best_path**.

for each DTW pair (R_{ix}, R_{iy}) and (X_{ix}, X_{iy}) in **best_path**:

$$\mathbf{x_cords} += [R_{ix} * \epsilon * h_{rs} + X_{ix} * (1 - \epsilon * h_{rs})]$$

$$\mathbf{y_cords} += [R_{iy} * \epsilon * h_{rs} + X_{iy} * (1 - \epsilon * h_{rs})]$$

Let $\mathbf{N}$ denote the vector of the new weights at the existing timestamps.

for each timestamp t in the existed timestamps of r :

if there is only one element $\mathbf{N}'_t$ in **x_cords** rounded to t

$$\mathbf{N} += [\mathbf{N}'_t]$$

if there is a list of (more than one) elements $\mathbf{N}'_t(s)$ in **x_cords** rounded to t

$$\mathbf{N} += [\text{the mean of } \mathbf{N}'_t(s)]$$

return $\mathbf{N}$

2.4 Data description and preprocessing

The Pediatric Research Using Integrated Sensor Monitoring Systems (PRISMS) Utah Informatics Platform Center conducted a panel study of 10 participants with asthma (ages 5-51, 4 children and 6 adults) in 7 households near Salt Lake City, Utah from April 2017 to April 2018. These 7 households included a total of 16 children ranging from younger than school age to late teens, some of whom participated in the study. In 2 of the households, a non-participant child was later diagnosed with asthma. PM_{2.5}, temperature, and relative humidity were measured both inside and outside of each house using two sensors located inside and one outside the home. The deployed sensors were a commercial Dylos Corporation particle counter, modified to include sensors for humidity and temperature, and to include w-fi communications. The conversion of particle counts to $\mu\text{g}/\text{m}^3$ follows rules suggested previously (Vercellino, Sleeth, Handy, Min, & Collingwood, 2018). Baseline questionnaires on home characteristics provided information on the presence of “scented products” (scented/unscented candles, incense, cedar, sage, room air fresheners, other scented household products) and the presence of indoor “combustion sources” (gas stove or range top; gas oven; gas or fuel furnace; gas or fuel hot water heater; propane, kerosene or gas space heater; gas fireplace or stove; wood/pellet-burning fireplace/stove; coal-burning stove; or other combustion source). Participants (or their guardians, for some child participants) from these households were asked to submit daily questionnaires about asthma symptoms and medication usage in the past 24 hours, including frequency of use of rescue medication (“How often did your child

use an Albuterol or Xopenex inhaler or receive a nebulized treatment in the last 24 hours?”). Daily questionnaires were submitted electronically before bedtime (typically ~8 p.m.).

We applied several data processing steps. To start, we used a median filter with kernel size of 3 to smooth the signals and remove extreme outliers. In some instances, more than one sensor was collecting data in a given microenvironment (either indoors or outdoors). To reduce the time-series from the two sensors into a single time-series, we calculated the mean of the minute-level concentrations. Future work might consider alternative methods (e.g., maximum or a single representative sensor). The time-series contained a considerable amount of missing data due to various reasons (e.g., sensors stopped working, malfunctions in the data transfer pipeline). We first truncated the data to remove time periods with consecutive missing values longer than 1 hour, and we then used a bidirectional linear interpolation method to fill-in periods of missing values with durations less than 1 hour. After that, we used a min-max scaler to rescale all pollutant values into a range from 0 to 1. Next, we used a non-overlapping 1-day windowing approach to separate the data into daily time-series (starting from 8 p.m. and ending at 8 p.m. on the next day). Finally, we excluded the incomplete days with less than 1,440 timestamps (the total number of minutes per day). Therefore, each daily observation has 6 time-series (temperature, humidity and PM_{2.5} indoors and outdoors) with length of 1,440. In this study, we only used outdoor temperature and indoor PM_{2.5} to test the proposed methodology. Matching days with exposure data to days with daily asthma questionnaires resulted in a total of 823 days of observations with complete data for both exposure time-series and questionnaires from 10 patients in 7 households.

2.5 Experiments to evaluate DTW-SOM

We designed and conducted 3 experiments to evaluate DTW-SOM and one additional experiment for examining the relationship between PM_{2.5} indoor sources and asthma control as defined by the use of an inhaler. We used the 24-hr outdoor temperature time series for the 3 evaluation experiments, and we used the 24-hr indoor PM_{2.5} time series for the knowledge discovery experiment. The evaluation of DTW-SOM is twofold: (1) the clustering performance based on intra-cluster similarity and purity; and (2) the spatial autocorrelation in the mapping layer based on Moran's I.

In the first experiment, we demonstrated the difference between the alignment and training rules in original SOM vs DTW-SOM by reporting the intermediate results of a single iteration of each when applied to a single exemplar input data sample (randomly selected) from the 24-hour outdoor temperature time series in our dataset. This experiment aimed to demonstrate the closeness between the trained weights and the initial weights under the two training rules and to compare their abilities for maintaining the patterns of the input time-series during training.

In the second experiment, under the assumption of a 10 by 10 output space (this parameter can be tuned in future applications), we compared results from: (a) the original SOM with Euclidean distance measurement, (b) the SOM with only DTW distance replacement, and (c) the DTW-SOM proposed in this

paper. We visually compared the output neuron’s weights generated by the three methods. The divergence of the three weights in each neuron indicates the performance of the three SOM methods.

In the third experiment, we compared the observations of each neuron between the original SOM and the DTW-SOM on clustering outdoor temperature. We also stratified the observations by seasons and compared the distributions and spatial autocorrelations of the observations from the two methods.

Finally, in the knowledge discovery experiment, we ran DTW-SOM on indoor PM_{2.5} and stratified the observations belonged to each neuron based on different PM_{2.5} sources and inhaler usage to gain insight into the potential association between PM_{2.5} sources and inhaler usage for asthma control.

3. Results

3.1 Comparing the training rules

A one-minute resolution input time series of measured ambient temperature (blue line) and its BMU (orange line), and the intermediate training results of the updated BMU after one iteration (green line) are shown in Figure 5. The red dashed lines delineate the alignment for updating the weights of the BMU. In both subplots, the orange BMU was determined by DTW distances. The updated BMU retains the clear peak observed in the input data under the DTW training rules used to update the BMU. However, the updated BMU appears much “smoother”, having lost the peak in the input data, under the Euclidean training. This plot only shows one iteration of training. After a large number of training iterations, the weights adjusted by the original SOM rule will largely lose the peak pattern even though the BMU at each iteration was picked by DTW distances (this conclusion is supported by the experiment in Section 3.2). We calculated and compared the Euclidean distances between the trained weights and the input data under both scenarios, as well as the variances of the trained weights, the initial weights and the input data to quantify the variation in the representative pattern extracted by the two training rules.

Table 2: Quantitative comparison of the trained weights resulting from one iteration of the Euclidean training rule versus the DTW training rule.

Summary statistic	Euclidean Alignment	DTW Alignment
Euclidean distance between the trained weight and the initial (BMU) weight	7.769	7.372
Variance of the trained weights*	0.0118	0.0137

*Variance of the input time-series was 0.0164 and the variance of the initial (best matching unit, BMU) weights was 0.0103.

Table 2 shows that the trained weights are still closer to the input time series under the DTW training rules than under Euclidean training rules, suggesting that the DTW trained weights better represent the input

data. The variance of the trained weights from the DTW training (0.0137) was closer to that of the original input time series (0.0164) than that of the Euclidean training (0.0118), indicating that DTW trained results better captured the patterns of the input data.

3.2 Comparing the diurnal patterns from the results

Because outdoor temperature is anticipated to exhibit a regular diurnal trend (i.e. peak temperature at midday), we used it to compare results from the original SOM, SOM with only DTW distance replacement, and DTW-SOM. Figure 6 shows the final weights of the neurons for the three methods.

The original SOM using Euclidean distance (the blue lines in Figure 6) produced a poor topological transformation of the diurnal patterns of outdoor temperature into the 2-D spatial relationship. For example, the upper left corner and the lower right corner should represent a very different diurnal pattern, since they are far away from each other on the 2-D map; however, the diurnal patterns they represent are quite similar, and this conflicts with the aim of SOM. SOM using a DTW distance measure (yellow lines) produced an output map that better retained the original topology relationship of the input data, but it also potentially over-smoothed the representative typical diurnal patterns. For example, the closer the neurons are to each other in the 2-D space, the more similar their weights (shapes) are for the yellow lines. There is also a gradual decrease in the daily variation from left to right and a gradual increase in the baseline temperature from top to bottom. Especially, in the lower right part of Figure 6, the yellow lines distinctly diverge from the blue lines. However, many of the yellow line neurons have “flattened” patterns, due to the Euclidean updating rules as demonstrated earlier. Only a few blue lines show flattened patterns (e.g., the middle right). In contrast, DTW-SOM (i.e., the red lines) both retains the original topology (neurons close to each other have more similar patterns than those far from each other) and maintains the shapes of the diurnal patterns, with little evidence of “flattened” patterns. The average Euclidean distances of each neuron from its eight Moore (nearest surrounding) neighbors was: 54.05 for the original SOM (blue), 48.89 for SOM with DTW distance (yellow), and 46.03 for DTW-SOM (red). This indicates that DTW-SOM produced much stronger neighborhood relationships than the original SOM.

3.3 Comparing the distributions of the observations belonged to each neuron

Using outdoor temperature, we compared the distributions of the observations belonged to each neuron between the original SOM and DTW-SOM. For the original SOM (Figure 7A), a large number of observations (22.23%, 183 of 823) matched to a single neuron in the upper right corner whereas for DTW-SOM (Figure 7C) more neurons had a mid-range number of observations, indicating these neurons were representative of a larger number of common diurnal patterns. When stratifying by the season of the year, we found that the gradual changes in patterns across the output map corresponded well with the expected seasonal differences in temperature for both methods. The observations in the lower left were mainly from summer and those in the upper right were from winter. DTW-SOM resulted in a slightly more distinct separation of observations from fall in the upper right than for the original SOM.

Table 3: Moran’s I, mean entropy, and mean Euclidean distances to centroids of the four seasons*

Season	Moran's I**		Mean Entropy (purity)		Mean Euclidean Distance to cluster centroid	
	Original SOM	DTW-SOM	Original SOM	DTW-SOM	Original SOM	DTW-SOM
Spring	0.325	0.373	4.965	4.907	5.889	5.862
Summer	0.416	0.616	5.594	5.313	7.924	7.881
Fall	0.269	0.464	4.532	4.107	6.451	5.949
Winter	0.604	0.510	5.071	4.985	7.153	6.776

* For each neuron, a season class label was determined by majority vote.

** Moran's I of both methods are all significant above the threshold $-1/(N-1) = -0.001$ for $N = 823$, indicating that both methods produced highly spatially autocorrelated season clusters.

The season-specific intra-cluster similarity was higher for DTW-SOM than for the original SOM (Table 3: larger mean Euclidean Distance to cluster centroids and higher purity measured by mean entropy). The spatial autocorrelation of the neurons within each season's cluster was also higher for DTW-SOM (Moran's I in Table 3). These results confirm that DTW-SOM better transformed the internal relationships of the input time series into spatial relationships on the 2-D maps.

3.4 Indoor PM_{2.5}, asthma inhaler usage, and household characteristics

We ran DTW-SOM on indoor PM_{2.5} time series and then stratified the observations by participant-level daily asthma inhaler usage and two different household-level variables on indicating the presence of potential indoor air pollution sources ("scented devices" and "combustion sources"). In Figure 8, we found that the days with higher indoor PM_{2.5} variabilities (primarily the neurons located at the top-right and the bottom-right corners of the map) appeared to also be the days when participants reported inhaler usage (subpanel B) and were also the days contributed by households with combustion sources (subpanel C) or scented devices (subpanel D). The planned future analyses would: (a) fine-tune the number of neurons in the DTW-SOM (here we arbitrarily chose a 10 x 10 map); and (b) evaluate the statistical evidence for associations between the closest matching typical diurnal pattern of indoor PM_{2.5} and health responses (e.g., daily inhaler usage or asthma symptoms).

Conclusions

This article developed and introduced DTW-SOM, a new time-series clustering method based on SOM. In applications of time-series clustering to identify diurnal exposure patterns, DTW-SOM better preserved the topology relationship of the input data and better summarized time-series patterns (e.g., retained diurnal temperature peaks) as compared to the original version of SOM. In our motivating application, we show that DTW-SOM can be applied to highly time-resolved measurements of environmental exposures to

discover and cluster typical diurnal patterns in temperature and PM_{2.5}. These diurnal patterns provide a richer profile of exposure than typical 24-hour, weekly, or monthly averages to assess health outcomes of interest. A limitation of the implementation of DTW-SOM presented here is that it supports analysis only on univariate time-series. Future extensions could include a multivariate version of DTW-SOM.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: NIH/NIBIB U24EB021996, U54EB021973, and U54EB022002.

Authors' contributions: KL introduced the idea, developed the algorithm and analyzed the data. KS conducted the 1-year PRISMS panel study, SPE and JPW was major contributors in writing the manuscript. All authors participated in the conceptual approach and revisions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: The authors gratefully acknowledge financial support from NIH/NIBIB U24EB021996, U54EB021973, and U54EB022002.

References

- Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. *Information Systems*, 53, 16-38. doi:<https://doi.org/10.1016/j.is.2015.04.007>
- Auchincloss, A. H., Roux, A. V. D., Dvorchak, J. T., Brown, P. L., Barr, R. G., Davignus, M. L., . . . O'Neill, M. S. (2008). Associations between recent exposure to ambient fine particulate matter and blood pressure in the Multi-Ethnic Study of Atherosclerosis (MESA). *Environmental Health Perspectives*, 116(4), 486-491. doi:10.1289/ehp.10899
- Berndt, D. J., & Clifford, J. (1994). *Using dynamic time warping to find patterns in time series*. Paper presented at the Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle, WA.
- Cherif, A., Cardot, H., & Bone, R. (2011). SOM time series clustering and prediction with recurrent neural networks. *Neurocomputing*, 74(11), 1936-1944. doi:10.1016/j.neucom.2010.11.026

- Cleasby, I. R., Wakefield, E. D., Morrissey, B. J., Bodey, T. W., Votier, S. C., Bearhop, S., & Hamer, K. C. (2019). Using time-series similarity measures to compare animal movement trajectories in ecology. *Behavioral Ecology and Sociobiology*, 73(11). doi:10.1007/s00265-019-2761-1
- Delfino, R. J., Zeiger, R. S., Seltzer, J. M., Street, D. H., & McLaren, C. E. (2002). Association of asthma symptoms with peak particulate air pollution and effect modification by anti-inflammatory medication use. *Environmental Health Perspectives*, 110(10), A607-A617. doi:10.1289/ehp.021100607
- Han, S. Q., Bian, H., Feng, Y. C., Liu, A. X., Li, X. J., Zeng, F., & Zhang, X. L. (2011). Analysis of the Relationship between O₃, NO and NO₂ in Tianjin, China. *Aerosol and Air Quality Research*, 11(2), 128-139. doi:10.4209/aaqr.2010.07.0055
- Juhasz, Z. (2009). *Motive Identification in 22 Folksong Corpora Using Dynamic Time Warping and Self Organizing Maps*. Paper presented at the ISMIR.
- Juhász, Z. (2007). ANALYSIS OF MELODY ROOTS IN HUNGARIAN FOLK MUSIC USING SELF-ORGANIZING MAPS WITH ADAPTIVELY WEIGHTED DYNAMIC TIME WARPING. *Applied Artificial Intelligence*, 21(1), 35-55. doi:10.1080/08839510600940116
- Kremer, H., Gunnemann, S., & Seidl, T. (2010, 13-13 Dec. 2010). *Detecting Climate Change in Multivariate Time Series Data by Novel Clustering and Cluster Tracing Techniques*. Paper presented at the 2010 IEEE International Conference on Data Mining Workshops.
- Liu, S. L., Krewski, D., Shi, Y. L., Chen, Y., & Burnett, R. T. (2003). Association between gaseous ambient air pollutants and adverse pregnancy outcomes in Vancouver, Canada. *Environmental Health Perspectives*, 111(14), 1773-1778. doi:10.1289/ehp.6251
- Mueen, A., Keogh, E., & Assoc Comp, M. (2016). Extracting Optimal Performance from Dynamic Time Warping. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2129-2130. doi:10.1145/2939672.2945383
- Niennattrakul, V., & Ratanamahatana, C. A. (2007, 26-28 April 2007). *On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping*. Paper presented at the 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07).
- Okada, S., & Hasegawa, O. (2008, 8-11 Dec. 2008). *Motion recognition based on Dynamic-Time Warping method with Self-Organizing Incremental Neural Network*. Paper presented at the 2008 19th International Conference on Pattern Recognition.
- Pearce, J. L., Waller, L. A., Chang, H. H., Klein, M., Mulholland, J. A., Sarnat, J. A., . . . Tolbert, P. E. (2014). Using self-organizing maps to develop ambient air quality classifications: a time series example. *Environmental health : a global access science source*, 13, 56-56. doi:10.1186/1476-069X-13-56

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61(4), 241-254. doi:10.1007/bf00203171

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5), 561-580.

Scepi, G., & Romano, E. (2006). *Integrating time alignment and Self Organizing Maps for Classifying Curves*.

Vercellino, R. J., Sleeth, D. K., Handy, R. G., Min, K. T., & Collingwood, S. C. (2018). Laboratory evaluation of a low-cost, real-time, aerosol multi-sensor. *Journal of Occupational and Environmental Hygiene*, 15(7), 559-567. doi:10.1080/15459624.2018.1468565

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11), 1857-1874. doi:<https://doi.org/10.1016/j.patcog.2005.01.025>

Wu, R., & Keogh, E. J. (2020). FastDTW is approximate and Generally Slower than the Algorithm it Approximates. *arXiv:2003.11246 [cs, stat]*.

Figures

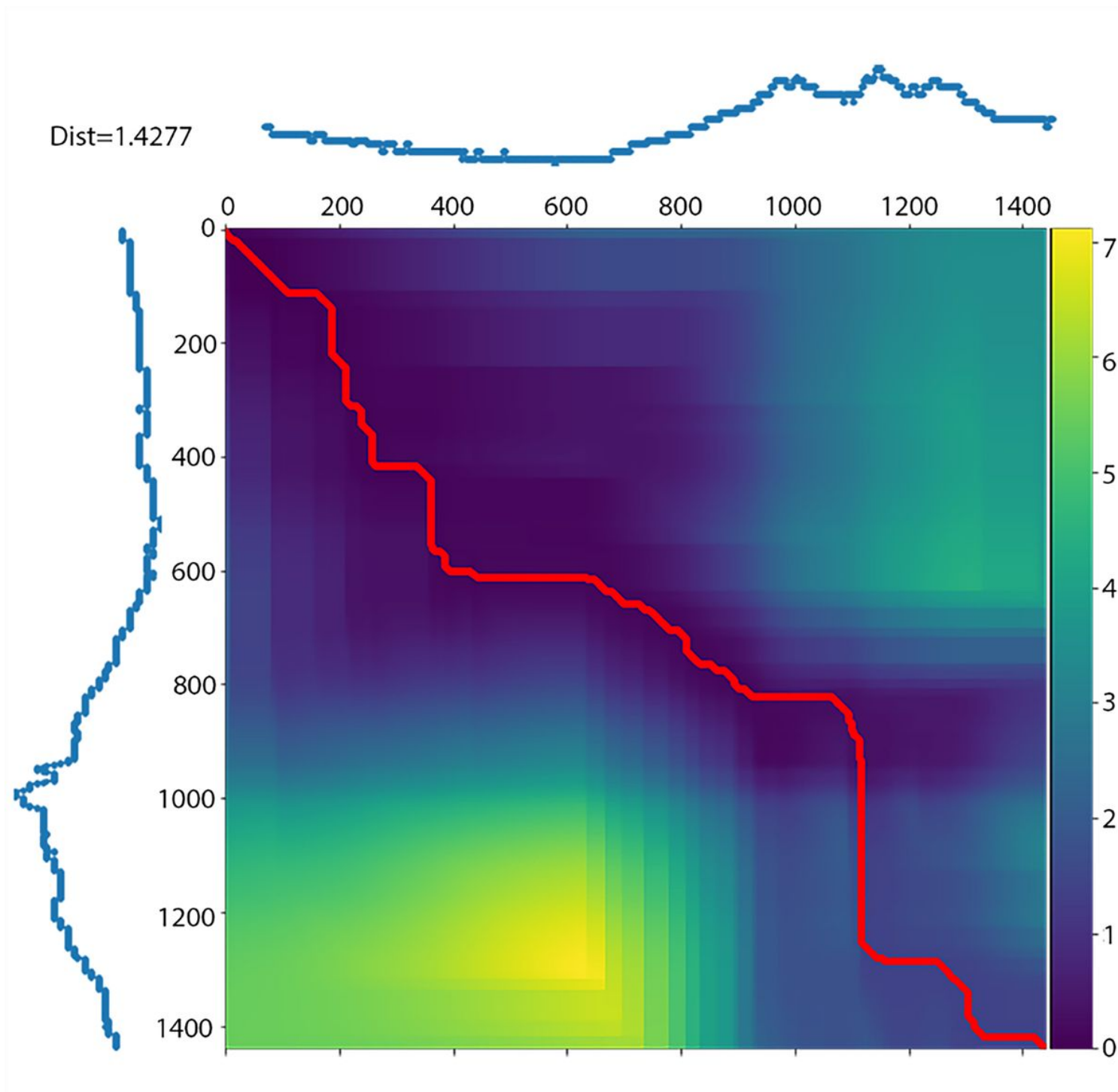


Figure 1

Example Dynamic Time Warping (DTW) alignment of two 24 hour time-series of outdoor temperature (at minute-level resolution so $60 \times 24 = 1,440$ minutes long) from two measurement sites, with the color gradient displaying all pairwise Euclidean distances between time points (blue indicates the shortest distances) and the red line shows the shortest path

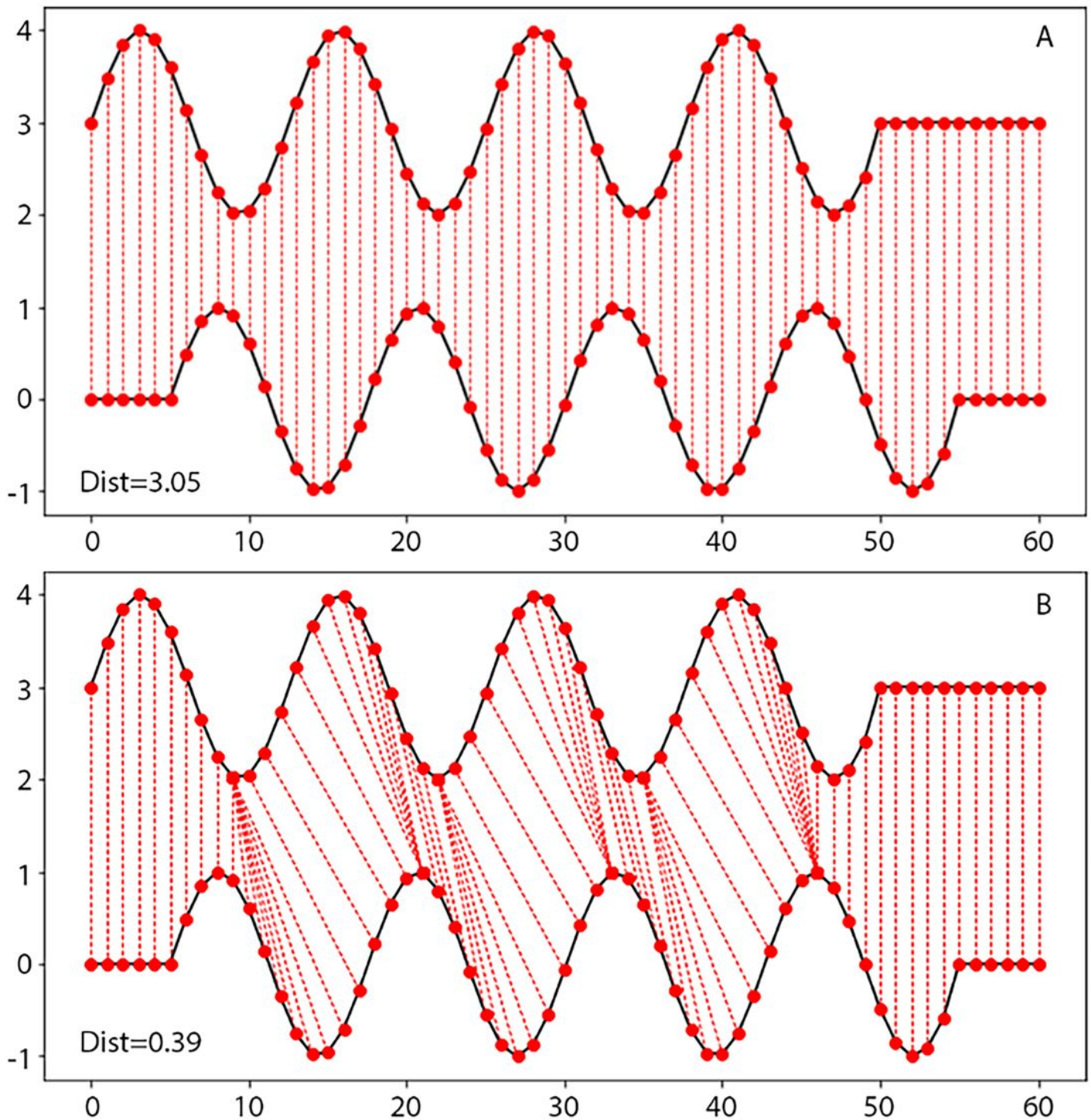


Figure 2

Euclidean (A) and DTW (B) alignment strategies applied to two synthetic signals with different baselines and a temporally misaligned internal sine curve pattern. Dist represents the component-wise distance between time series.

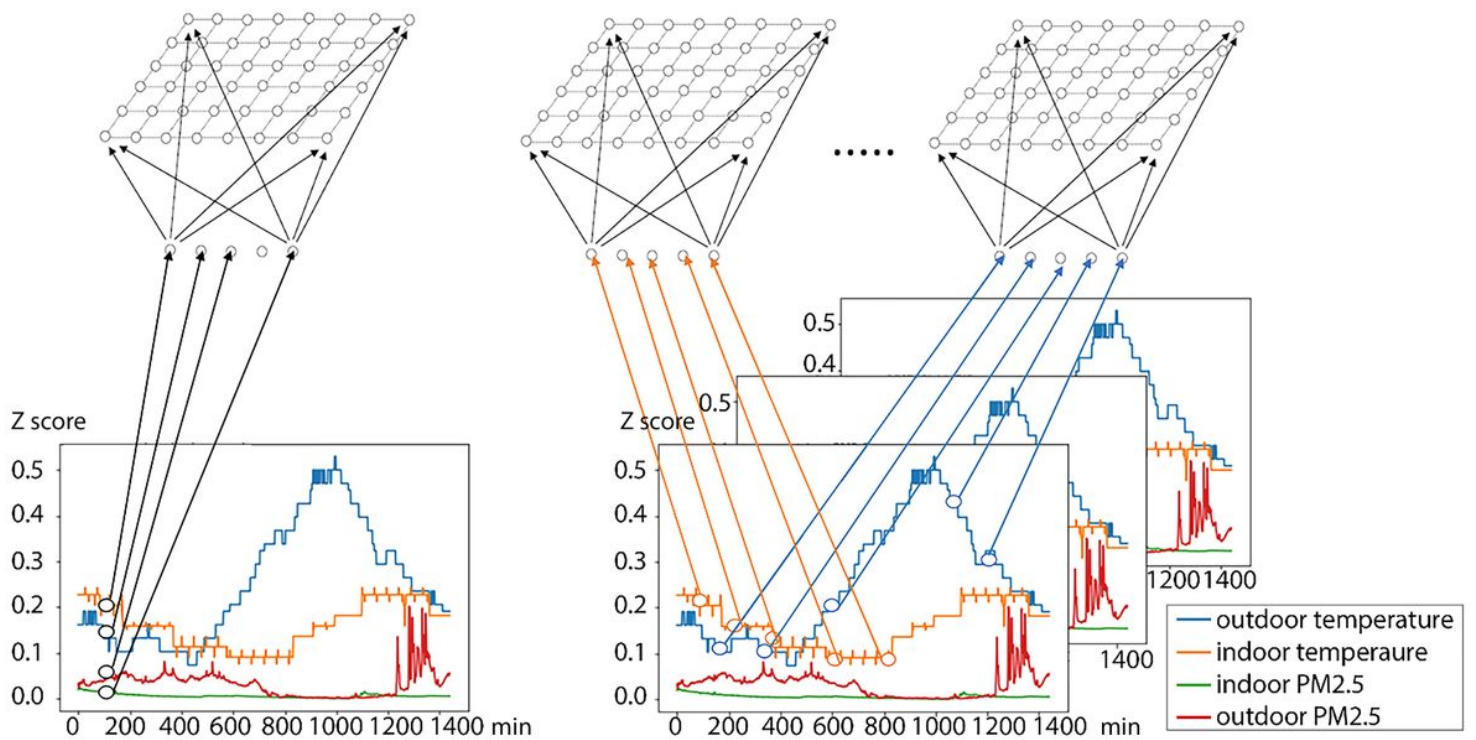


Figure 3

Conceptual diagrams contrasting the use of SOM to: (a) discover multipollutant patterns, as in Pierce et al. (2014), and (b) diurnal patterns of a single pollutant, as proposed in this study

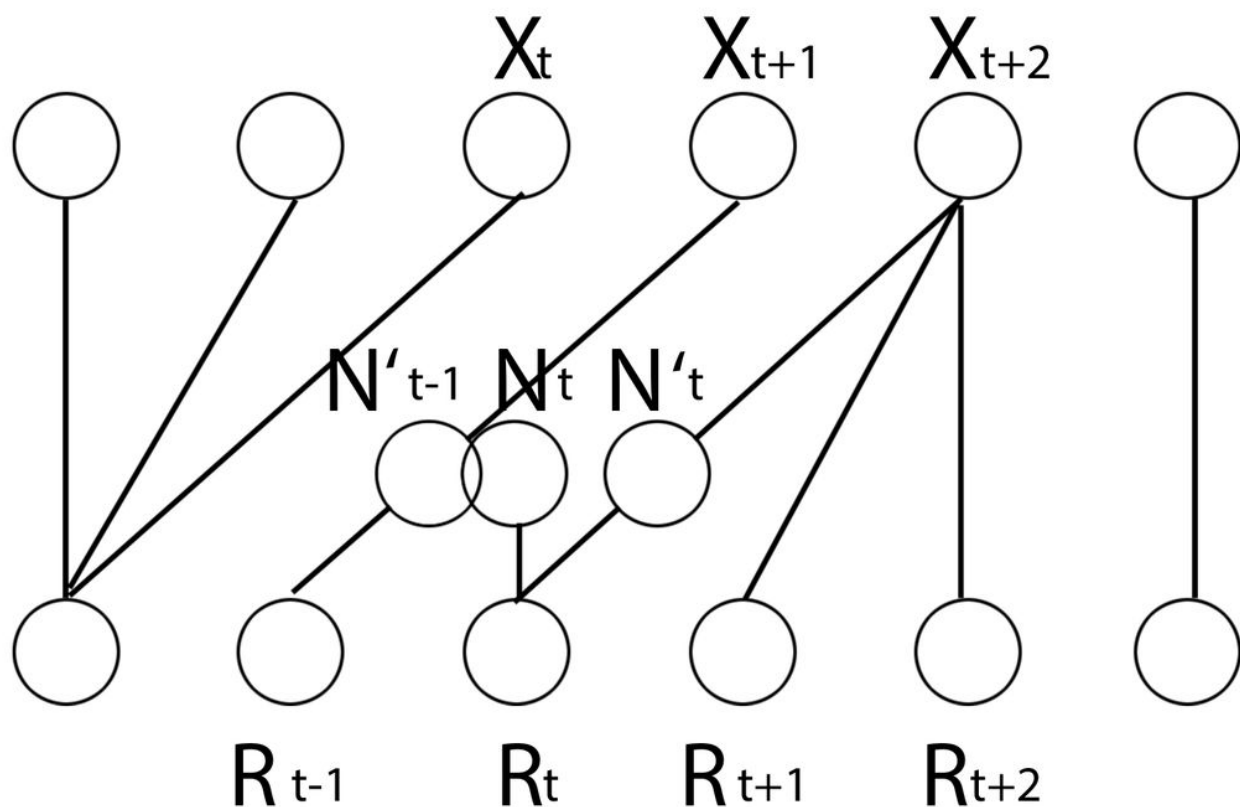


Figure 4

Calculation of new neuron weights N_t using DTW alignment. N'_t and N'_{t-1} are the weighted means of two adjacent DTW pairs; however, neither of them falls into any existing timestamps. In order to calculate N_t , we need to approximate its value using N'_t and N'_{t-1}

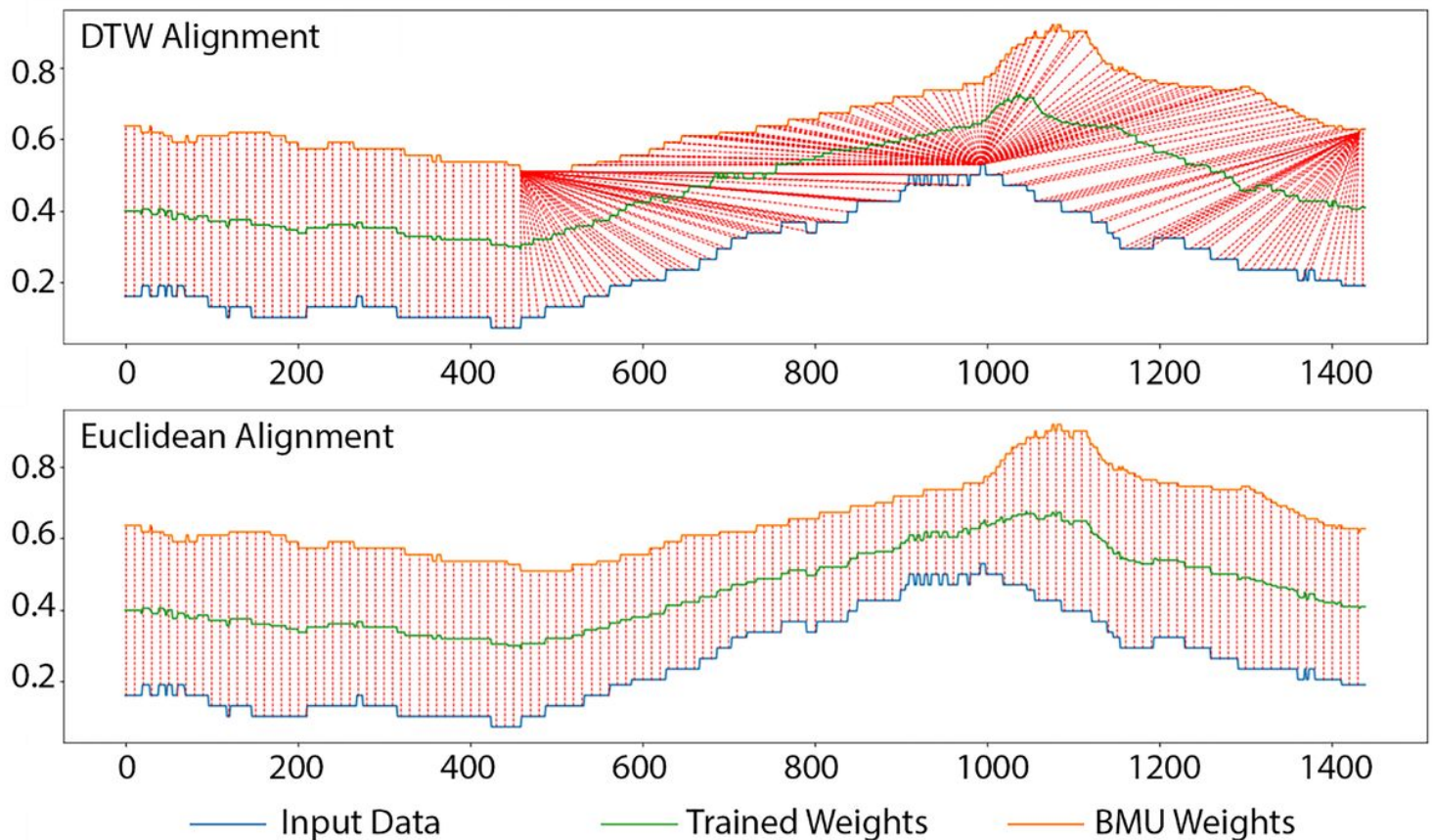


Figure 5

Visual comparison of the trained weights resulting from one iteration of the Euclidean training rule versus the DTW training rule, where input data refers to the original time-series, trained weights refers to the weights after the iteration, and BMU weights refers to the weights at the previous iteration

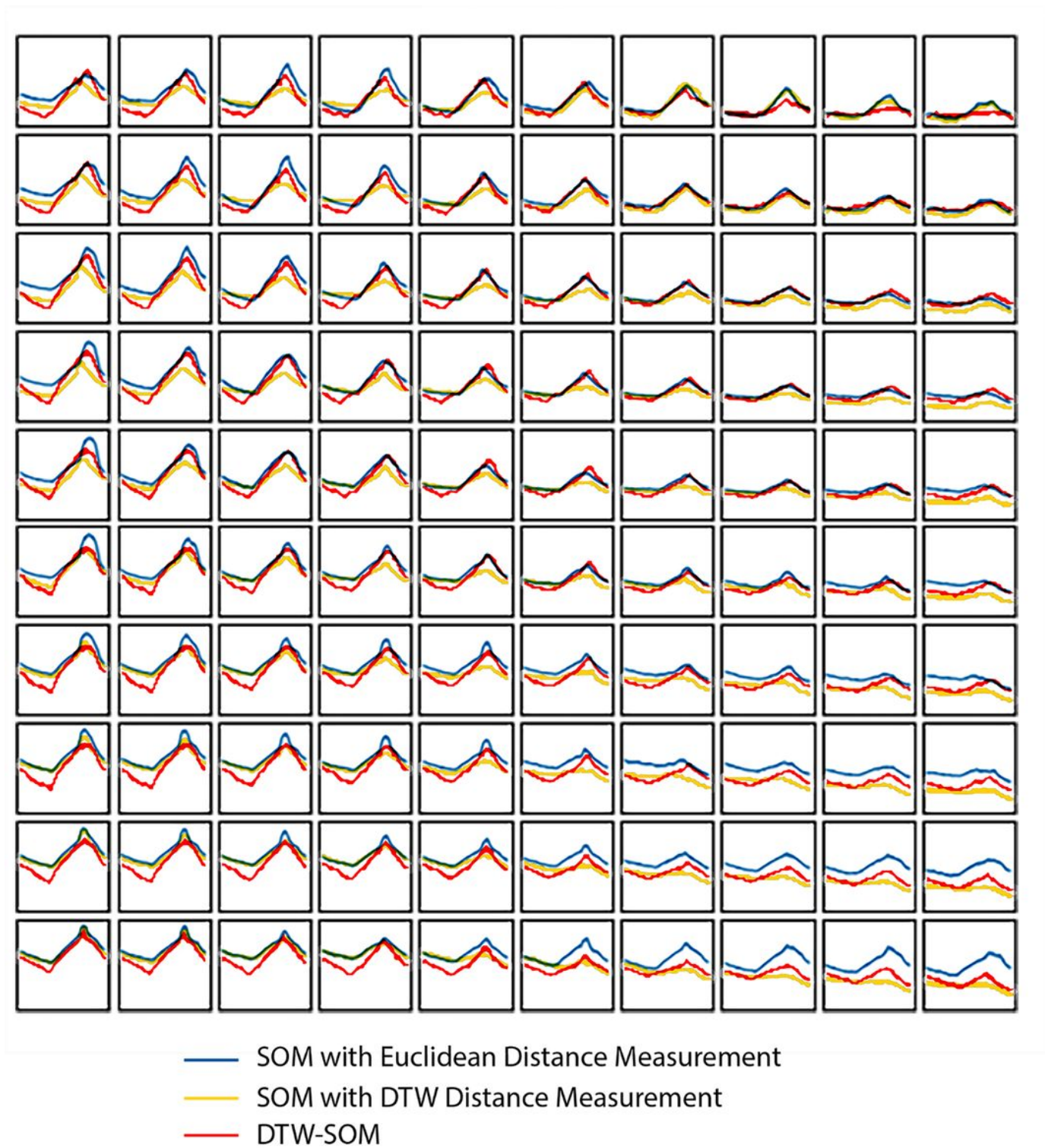


Figure 6

Comparing the results of the original SOM (blue), SOM with only DTW measurement (yellow), and DTW-SOM (red) on outdoor temperature. Each cell represents a neuron, and the curved line represents its final weights

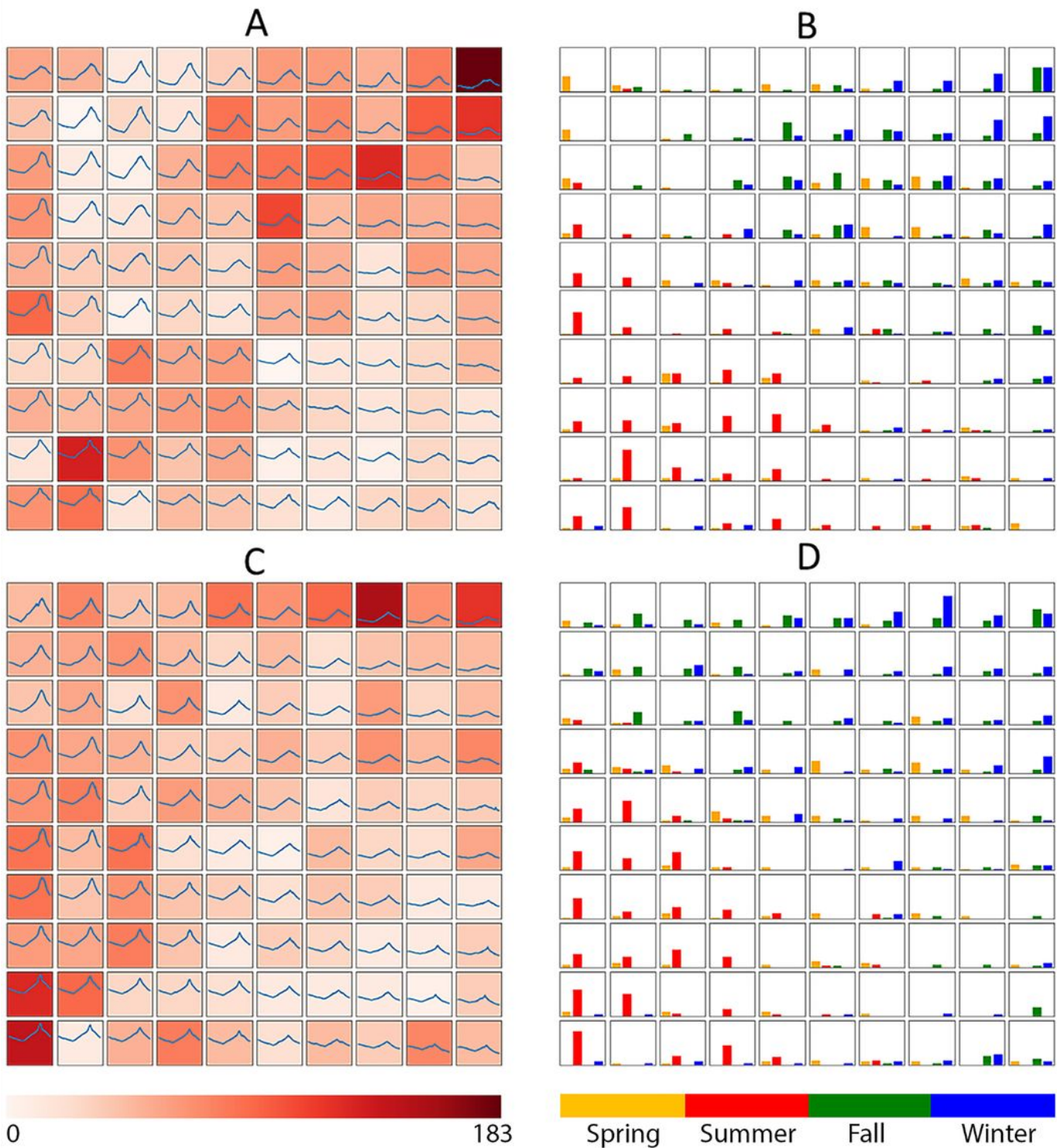


Figure 7

Distribution of the number of observations per neuron overall (A, C) and by season (B, D) for the outdoor temperature time-series using the original SOM (A, B) and DTW-SOM (C, D). The darker red color in A and C indicates a larger number of observations and bar plots in B and D indicate the number of observations in each season

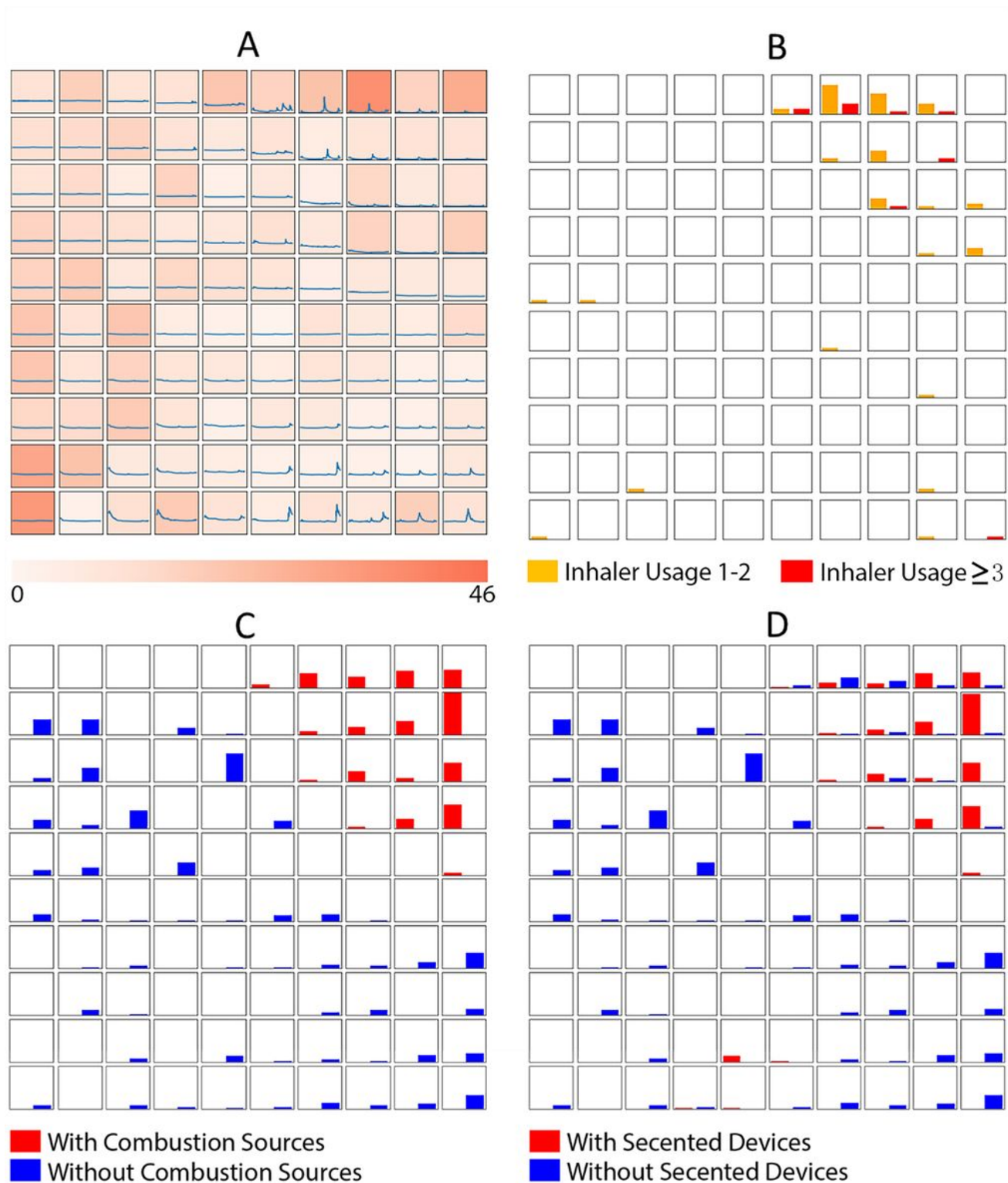


Figure 8

Stratification of the observations per neuron (indoor PM_{2.5}). A: the diurnal patterns represented by each neuron. B: daily inhaler usage frequency (orange: 1-2, red: ≥ 3). C: existence of combustion sources (red: yes, blue: no).* D: existence of scented devices (red: yes, blue: no).** * Combustion source: Gas stove or range top, Gas oven, Gas or fuel furnace, Gas/fuel hot water heater, Gas, Propane, kerosene, Gas space heater, Gas fireplace or stove, Wood/pellet-burning fireplace/stove, Coal-burning stove, and other

combustion source ** Scented device: Scented candles, Unscented candles, Incense, Cedar, Sage, Room air fresheners (i.e. plug ins, stick ups) and other scented household products