Identifying strongly correlated groups of sections in a large motorway network

Shanshan Wang *, Michael Schreckenberg and Thomas Guhr

Faculty of Physics, University of Duisburg–Essen, Duisburg, Germany

February 16, 2022

Abstract. In a motorway network, correlations between the different links, i.e. between the parts of (different) motorways, are of considerable interest. Knowledge of fluxes and velocities on individual motorways is not sufficient, rather, their correlations determine or reflect, respectively, the functionality of and the dynamics on the network as a whole. These correlations are time dependent as the dynamics on the network is highly non-stationary, as it strongly varies during the day and over the week. Correlations are indispensable to detect risks of failure in a traffic network. Discovery of alternative routes less correlated with the vulnerable ones helps to make the traffic network robust and to avoid a collapse. Hence, the identification of, especially, groups of strongly correlated road sections is needed. To this end, we employ an optimized \(k\)-means clustering method. A major ingredient is the spectral information of certain correlation matrices in which the leading collective motion of the network has been removed. We identify strongly correlated groups of sections in the large motorway network of North Rhine-Westphalia (NRW), Germany. The groups classify the motorway sections in terms of spectral and geographic features as well as of traffic phases during different time periods. The representation and visualization of the groups on the real topology, i.e. on the road map, provides new results on the dynamics on the motorway network. Our approach is very general and can also be applied to other correlated complex systems.

Keywords: complex system, traffic network, correlation matrix, spectral decomposition, \(k\)-means clustering, strongly correlated groups

Contents

1 Introduction 2
2 Datasets 3
3 Methods 3
3.1 Standard correlation matrices 3
3.2 Reduced-rank covariance and correlation matrices 5
3.3 Eigenvalues of covariance and correlation matrices 6
3.4 Clustering with the principal eigenvectors 7
4 Empirical results 8
4.1 Identification of strongly correlated groups 9
4.2 Spectral features of strongly correlated groups 9
4.3 Geographic features of strongly correlated groups 13

*shanshan.wang@uni-due.de
1 Introduction

A traffic network as a complex system [1,2] is composed of all road sections on all links, i.e. parts of the different motorways between ramps and crosses. These sections are connected with each other via the motion of the vehicles, resulting in highly non-stationary time series of traffic flows and velocities. Similar behavior of these time series is revealed in correlations of traffic flows or velocities between different road sections [3]. We notice that these can be, for example, neighboring sections on the same motorway or sections on different motorways far away from each other. Obviously, the correlations of the latter are of particular interest. Here, we recall that correlations do not necessarily reflect causalities, nevertheless the correlations are needed to assess the functionality of the network. The identifications of groups with particularly strong correlations is of paramount importance. This concept is of course rather general and relates our study to those of many other complex systems with correlated components. Local perturbations, such as blackouts in power grids [4], congestions in traffic networks [5,6] and bankruptcies in credit networks [7] can prompt cascading failures of the whole system. Correlations between road sections thus reveal or cause, respectively, risks of failures in the traffic network as a whole. Detection of alternative routes is called for to effectively reduce the traveling time and enhance the traffic efficiency, but also to make the traffic network more robust against a collapse. In practice, determining alternative routes, however, is costly due to an often high amount of connections from origin to destination. To know a priori the strongly correlated groups of sections considerably facilitates the detection of alternative routes. Apart from the above issue, identifying such strongly correlated groups is also highly relevant for discovering critical bottlenecks [8], exploring critical phenomena [9], recognizing metastable or quasi-stationary states [3,10] and for other aspects in traffic networks.

Our goal is the identification of strongly correlated groups of sections in the large motorway network of North Rhine-Westphalia (NRW), Germany. We transfer and extend methods previously developed for correlated financial markets. In this context, the industrial sectors are the strongly correlated groups that we aim to identify here. Stocks within the same industrial sector show strong correlations indicating collective motion within this sector, but not across different sectors. They reveal themselves in the spectrum of the correlation matrix for the entire market [11–13]. More precisely, one outlier, i.e. an eigenvalue outside the bulk, corresponds to each industrial sector. These outliers are fairly robust, even though financial markets are highly non-stationary. The largest eigenvalue, however, belongs to the market as a whole, indicating the collective motion of the entire system, comprising all stocks in the market. This collectivity overspreads the above collectivity in the sectors. Hence, a method is called for to assess the sector collectivity in the moving frame of the market collectivity. Put differently, we need to remove the market collectivity to obtain a much clearer view on the dynamics and the interaction of the different industrial sectors. Such a method has recently been developed [3,14,15], it is based on a proper rank-reduction of the correlation matrix. The spectrum of this reduced-rank correlation matrix is then used to better study the industrial sectors. With some necessary modifications and with the help of a $k$-means clustering method, we transfer this approach to our traffic data to identify strongly correlated groups of motorway sections.

There is an important difference between the analysis of a financial market and traffic data. Motorway sections have a location in the traffic network, i.e. there is a spatial aspect which has no analog in a stock market. A network between these stocks derives from the correlations only, and such a network also exists for the road sections, but this is a virtual one, while the true geographic
Identifying strongly correlated groups of sections in a large motorway network
topology of the motorways forms a real network whose functionality is the ultimate objective of our study.

We also notice that our study is different from previous ones [8–10] focusing on urban networks. Our motorway network covers urban traffic and traffic in the countryside, thereby adding a new aspect highly relevant for territorial states.

The paper is organized as follows. In section 2, we introduce our data set. In section 3, we briefly sketch the construction of reduced-rank correlation matrices, compare the spectral information of the standard and reduced-rank correlation matrices, and develop an optimized \( k \)-means clustering method. In section 4, we identify strongly correlated groups of motorway sections with empirical data, disclose their spectral and geographic features, and further figure out the relation between the dominant eigenvalues and geographic distributions of sections. We finally conclude our results in section 5.

2 Datasets

Our traffic data are collected by inductive loop detectors from \( N = 1179 \) sections on 22 motorways in North Rhine-Westphalia (NRW), Germany, shown in figure 1. The data with resolution of one minute covers 80 discontinuous days, including 64 workdays and 16 holidays, in 2017. Here the weekends and public holidays of NRW in 2017 are all named as holidays. On each selected day, the ratio of missing values in the traffic data for each section is less than 60%. To be more specific, if considering the traffic data of each section at each day as a sub-dataset, 97.95%, 92.44% and 59.65% sub-datasets have less than 20%, 10% and 1% missing values, respectively. Regarding the data quality, the missing values are filled by their nearest non-missing values. We verified that the way of filling missing values has a negligible effect on our empirical results. This is important, as the missing values have a negative effect on the following spectrum decomposition. For each time, the data set includes the information of traffic flows and velocities for every lane on every section. The traffic flow gives the number of vehicles per unit time. Divided by the velocity, it yields the flow density that measures the number of vehicles per unit distance. For each section \( n \) at each time \( t \), we aggregate the traffic flows \( q_{nt} \) and the flow densities \( \rho_{nt} \) across all lanes \( l \). The average velocity \( v_n(t) \) for the same section at the same time can be obtained by

\[
v_n(t) = \frac{\sum_l q_{nt}}{\sum_l \rho_{nt}}, \quad n = 1, \cdots, N.
\]

Our study considers the case of all vehicles without distinguishing cars and trucks unless specific instructions, but distinguishes the cases of workdays and holidays due to different traffic behaviors.

3 Methods

To describe the methods for data analysis in detail, we first give the definition of the standard correlation matrix in section 3.1. To remove the effect of the largest eigenvalue from the standard correlation matrix, we derive a reduced-rank correlation matrix in section 3.2. We compare the spectral information between the standard and reduced-rank correlation matrices in section 3.3. With the spectral information of the reduced-rank correlation matrix, we then carry out the \( k \)-means clustering and further optimize the clustering with the help of silhouette values in section 3.4. The above spectral approach finally results in our strongly correlated groups.

3.1 Standard correlation matrices

For each section \( n \), we have a time series of velocities \( v_n(t) \) with the length \( T \). The mean value and the standard deviation of this time series can be expressed by

\[
\mu_n = \frac{1}{T} \sum_{t=1}^{T} v_n(t), \quad n = 1, \cdots, N,
\]

3
Identifying strongly correlated groups of sections in a large motorway network

all used sections toward
- the north-east
- the south-west
- motorways in NRW
- NRW state boundary

Figure 1: The geographic distributions of 1179 available sections on 22 motorways of NRW in Germany. The green background indicates the population density of districts in NRW. The darker the green background, the higher the population density is. The data of administrative borders of districts (green lines) in NRW, licensed under BY-2.0, is provided by © GeoBasis-DE / BKG 2020 [16, 17] and the data of population density in NRW, also licensed under BY-2.0, is provided by © Statistische Ämter des Bundes und der Länder, Germany [17, 18]. The data of motorways (black lines) and the outside administrative boundaries of NRW (grey lines), licensed under ODbL v1.0, is provided by © OpenStreetMap contributors [19, 20]. The map is developed with QGIS 3.4 [21].
Identifying strongly correlated groups of sections in a large motorway network

\[ \sigma_n = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (v_n(t) - \mu_n)^2}, \quad n = 1, \ldots, N, \] (3)

respectively. We normalize each element of this time series to zero mean and unit standard deviation by

\[ M_n(t) = \frac{v_n(t) - \mu_n}{\sigma_n}, \quad n = 1, \ldots, N. \] (4)

Thus, we obtain a \( N \times T \) data matrix \( M \) whose \( n \)-th row is the normalized time series \( M_n(t) \), \( t = 1, \ldots, T \). Therefore, the \( N \times N \) correlation matrix of sections is given by

\[ C = \frac{1}{T} M M^\dagger. \] (5)

As explained in reference [15], the largest eigenvalue of the correlation matrix \( C \) captures the collective behavior of significant sections in the whole motorway network in NRW. For example, the rush hours contribute to this collectivity, but they are not the only reason for it, as the temporal analysis in reference [15] reveals. Collective behavior present in financial markets is identified similarly. The largest eigenvalue [13, 22] is proportional to the average of all elements in the correlation matrix [13, 23]. We notice that the systems in question, financial markets as well as traffic networks, are highly non-stationary, featuring different dynamics and correlation structures at different times. Hence, the dynamics of the largest eigenvalue may be viewed as a moving frame from which we now wish to assess the remaining dynamics. Thus, to separate the group behavior from the collective behavior, we need to remove the effect of the largest eigenvalue from the correlation matrix.

One possible method is a linear regression with empirical data to obtain the residuals as the new data [12, 13], which yields a new correlation matrix without the effect of the largest eigenvalue from the standard correlation matrix. Here we remove the effect of the largest eigenvalue by combining a singular value decomposition with the reconstruction of correlation matrices [14]. The resulting correlation matrix is referred to as a reduced-rank correlation matrix, see reference [24] in another context.

### 3.2 Reduced-rank covariance and correlation matrices

We normalize each element of time series of \( v_n(t) \) only to zero mean instead of to both zero mean and unit standard deviation

\[ A_n(t) = v_n(t) - \mu_n. \] (6)

The \( N \) time series \( A_n(t) \) form a new \( N \times T \) data matrix \( A \). We apply a singular value decomposition and expand \( A \) in a sum of dyadic matrices,

\[ A = \sum_{n=1}^{L} S_n U_n V_n^\dagger, \] (7)

where \( L = \min(N, T) \) and the \( S_n \) are the singular values. There are \( N \) singular values for \( N < T \) and \( T \) for \( T \leq N \). Furthermore, \( U_n \) and \( V_n \) are the corresponding left and right singular eigenvectors, respectively, where \( U_n \) has \( N \) and \( V_n \) has \( T \) components. Removing the largest eigenvalue from \( A \) yields a new \( N \times T \) data matrix,

\[ \tilde{A} = \sum_{n=1}^{L-1} S_n U_n V_n^\dagger. \] (8)

We introduce a \( T \) dimensional unit column vector \( e = (1, \ldots, 1) \) and a \( P \) dimensional zero column vector \( \emptyset_P = (0, \ldots, 0) \) with \( P = N \) or \( T \) such that

\[ A e = \emptyset_N \quad \text{and} \quad A^\dagger A e = \emptyset_T. \] (9)
The first of equation (9) is simply the normalization of all time series \(A_n(t), \ t = 1, \ldots, T\) to zero means, written in a linear-algebra notation. From equations (7) and (9) and due to the linear independence of the \(U_n\), we find

\[ V_n^t e = 0, \quad (10) \]

for all \(n\) in the case of \(N < T\) due to the existence of \(N\) non-zero singular values. In the case of \(T \leq N\), there are \(T - 1\) non-zero singular values and one zero singular value, such that equation (10) is fulfilled for all \(n\) except for the one when \(S_n = 0\). In any case, the following equation holds,

\[ \bar{A} e = \sum_{n=1}^{L-1} S_n U_n V_n^t e = \varnothing_N. \quad (11) \]

Hence, all time series in \(\bar{A}\) are normalized to zero mean. The reduced-rank data matrix \(\bar{A}\) therefore yields a well-defined covariance matrix,

\[ \Sigma = \frac{1}{T} \bar{A} \bar{A}^t, \quad (12) \]

named reduced-rank covariance matrix. The elements of each row in \(\bar{A}\) can be normalized to unit standard deviation by dividing out the standard deviation of the row,

\[ \tilde{\sigma} = \frac{1}{T} \bar{\sigma} \]

where \(\bar{\sigma}\) is the diagonal matrix of the standard deviations

\[ \bar{\sigma} = \text{diag}(\bar{\sigma}_1, \ldots, \bar{\sigma}_N) \quad \text{with} \quad \bar{\sigma}_n = \sqrt{\Sigma_{nn}}. \quad (14) \]

Utilizing the definition of a correlation matrix, we find the \(N \times N\) reduced-rank correlation matrix

\[ \tilde{\mathcal{C}} = \frac{1}{T} \tilde{\sigma} \tilde{\sigma}^t = \bar{\sigma}^{-1} \Sigma \bar{\sigma}^{-1}. \quad (15) \]

For more details on the reduced-rank correlation matrix, we refer the reader to reference [14].

### 3.3 Eigenvalues of covariance and correlation matrices

In the following, we compare the spectral information between the standard and reduced-rank correlation matrices. The spectral decomposition of the standard covariance matrix reads

\[ \Sigma = \frac{1}{T} AA^t = \sum_{n=1}^L \Lambda_n U_n U_n^t, \quad (16) \]

where the eigenvalues

\[ \Lambda_n = \frac{S_n^2}{T} \quad (17) \]

are directly related to the singular values. The standard covariance and correlation matrices \(\Sigma\) and \(\mathcal{C}\), respectively, are related by

\[ \Sigma = \sigma \mathcal{C} \sigma, \quad \sigma = \text{diag}(\sigma_1, \ldots, \sigma_L) \quad (18) \]

with the standard deviations \(\sigma_l = \sqrt{\Sigma_{ll}}\). With equations (17) and (18), we are able to expand equation (15) as

\[ \tilde{\mathcal{C}} = \bar{\sigma}^{-1} \left( \Sigma - \Lambda_l U_l U_l^t \right) \bar{\sigma}^{-1} = \bar{\sigma}^{-1} \left( \sigma \mathcal{C} \sigma - \frac{S_l^2}{T} U_l U_l^t \right) \bar{\sigma}^{-1}. \quad (19) \]

It is worth mentioning that there are in total \(N\) eigenvalues either for \(\Sigma\) or for \(\mathcal{C}\) due to the matrix dimensions \(N \times N\). The number of their non-zero eigenvalues is \(L\) in the case of \(N < T\) and \(L - 1\) in
Identifying strongly correlated groups of sections in a large motorway network

the case of $T \leq N$ [14, 25] with $L = \min(N, T)$. To avoid confusion, we list the numbers of non-zero eigenvalues in Table 1. In an ascending order, $\Lambda_L$ and $S_L$ are the largest non-zero eigenvalue and the largest non-zero singular values of $\Sigma$ and $A$, respectively. We further decompose the two correlation matrices in the above equation by

$$C = \sum_{n=1}^{L} \lambda_n u_n u_n^\dagger$$ (20)

and

$$\tilde{C} = \sum_{n=1}^{L} \tilde{\lambda}_n \tilde{u}_n \tilde{u}_n^\dagger ,$$ (21)

where we only consider the largest $L$ eigenvalues $\lambda_n$ and $\tilde{\lambda}_n$ of $C$ and $\tilde{C}$, respectively, and ignore the other zero eigenvalues. Besides, $u_n$ and $\tilde{u}_n$ are the corresponding $N$-component eigenvectors of $C$ and $\tilde{C}$, respectively. As there are $N - 1$ non-zero eigenvalues for $N < T$ and $T - 2$ non-zero eigenvalues for $T \leq N$ as shown in Table 1, the minimal eigenvalue is always zero, i.e., $\tilde{\lambda}_1 = 0$.

### 3.4 Clustering with the principal eigenvectors

Equation (21) reads in matrix form

$$\tilde{C} = \tilde{u} \tilde{\lambda} \tilde{u}^\dagger , \quad \tilde{\lambda} = \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)$$ (22)

with the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ in an ascending order, and $\tilde{u}$ is a $N \times N$ orthogonal matrix whose columns are the corresponding eigenvectors $\tilde{u}_n$, such that

$$\tilde{u} = [\tilde{u}_1 \cdots \tilde{u}_N] .$$ (23)

Either for $N < T$ or for $T \leq N$, the matrix $\tilde{C}$ always has $N$ eigenvalues in total, as listed in Table 1, and the $N$ corresponding eigenvectors. To reduce the noise and to lower the dimension to the relevant one for clustering, we focus on the large eigenvalues. To find $k$ correlated groups of sections, we use the eigenvector information corresponding to the largest $k - 1$ eigenvalues [26] and define the $N \times (k - 1)$ eigenvector matrix

$$\tilde{u}^{(k-1)} = [\tilde{u}_{N-(k-1)+1} \cdots \tilde{u}_N]$$ (24)

as our data for $k$-means clustering [27, 28]. It then follows that $k$ is the number of clusters (or groups) in $k$-means clustering.

To determine the number $k$, we resort to the Marchenko-Pastur eigenvalue density [29] as a qualitative guideline. The Marchenko-Pastur distribution, resulting from the spectral density of a correlation matrix,

$$\rho(\lambda) = \sum_{n=1}^{N} \delta(\lambda - \lambda_n) ,$$ (25)

is the large-$N$ eigenvalue density for a fully random correlation matrix. It is known [13, 30, 31] that it also describes the bulk of many large non-random correlation matrices. The bulk of eigenvalues is between $\lambda_-$ and $\lambda_+$

$$\lambda_\pm = 1 + \frac{N}{T} \pm 2\sqrt{\frac{N}{T}}$$ (26)

| matrix dimensions | $N < T$ | $T \leq N$ |
|-------------------|---------|------------|
| $A$ | $N \times T$ | $N$ | $T$ |
| $\Sigma$ | $N \times N$ | $N$ | $N$ |
| $C$ | $N \times N$ | $N$ | $N$ |
| $\tilde{C}$ | $N \times N$ | $N - 1$ | $T - 2$ |
with $T \neq N$. Typically, there are large eigenvalues outside the bulk, often way outside the bulk. They indicate strongly correlated groups [12, 13], and their number $k-1$ is the one we use for the clustering.

To carry out the clustering, we consider the $n$-th row of the matrix $\tilde{u}^{(k-1)}$, i.e., a vector, as an observation corresponding to section $n$. Therefore, clustering all these observations means clustering our sections. The components in this vector are the features for comparing the similarity of two observations. Here, we define the distance between two observations $i$ and $j$ as the squared Euclidean distance,

$$d_{ij} = \sum_{m=N-k+2}^{N} (\tilde{u}_{im} - \tilde{u}_{jm})^2,$$  

which entries in a distance matrix $d$. Employing the distance matrix, we implement $k$-means clustering [27, 28] for our observations. The $k$-means clustering mainly contains the following steps:

(a) Select $k$ initial centroids for observations.
(b) Compute distances from each observation to every centroid.
(c) Assign each observation to the cluster with the closest centroid.
(d) Recalculate the average of the observations assigned to each cluster in order to find a new centroid for each cluster.
(e) Repeat steps (b)–(d) until the assignments of observations do not change or iterations reach the preset maximal number.

The silhouette value quantifies how similar an observation is to its own cluster as compared to other clusters [32]. It ranges from -1 to +1, where a high positive value indicates an appropriate classification of an observation under its own cluster, while a low or negative value indicates a poor clustering configuration. To optimize our clustering, we carry out the whole procedure of clustering as follows:

(1) Perform $k$-means clustering with the squared Euclidean distance.
(2) Refine the clustering as follows:

(2.1) Validate the consistency within clusters by silhouette values and reassign all observations with negative silhouette values to an additional cluster, i.e., $(k+1)$-th cluster.
(2.2) Reassign each observation in $(k+1)$-th cluster separately to all $k+1$ clusters and calculate the silhouette value of that observation for every assignment.
(2.3) Reassign that observation to the cluster in which the observation acquires the maximal silhouette value comparing to in other clusters.
(2.4) Calculate silhouette values of all observations and reassign the observations with negative silhouette values to $(k+1)$-th cluster.
(2.5) Repeat steps (2.2)–(2.4) until the assignments of observations do not change or iterations reach the preset maximal number.

(3) Validate the consistency within clusters by silhouette values.
(4) Reorder the indices of $k+1$ clusters according to the contribution of eigenvectors.

4 Empirical results

We analyze empirical data of motorway sections in NRW, Germany. We apply the spectral approach to our empirical data and identify strongly correlated groups of motorway sections in section 4.1. We then identify spectral features and figure out dominant eigenvalues for each group in section 4.2. In section 4.3, we visualize and analyze geographic features of groups in terms of section-concentrated regions and traffic phases [33], including free and congested phases. We associate dominant eigenvalues with geographic distributions of motorway sections in section 4.3.
4.1 Identification of strongly correlated groups

We work out the $1179 \times 1179$ reduced-rank correlation matrices for workdays and holidays with the method described in section 3.2, where $N = 1179$ and $T = 1440$. The largest eigenvalue of the standard correlation matrix, i.e., $\lambda_{\text{max}} = 135.6$ for workdays and $\lambda_{\text{max}} = 112.8$ for holidays, respectively, are subtracted applying the described procedure. According to our previous experience, the resulting reduced-rank correlation matrix is free of collective behavior that affects the system as a whole, see references [14] for financial markets and [15] for the NRW motorway network. In the case of workdays, however, there is a deviation from the mentioned previous analyses. The numerical value $\lambda_{2\text{nd max}} = 131.3$ of the second largest eigenvalue is rather close to that of the largest. Usually, the numerical value of the second largest eigenvalue is considerably smaller. Obviously, the second largest eigenvalue in the case in question reflects the presence of another collectivity affecting a large part of the system as well. Nevertheless, the distribution of the eigenvector components corresponding to the largest and the second largest eigenvalue worked out in reference [15] show differences for the significant participants, i.e., the significant motorway sections. As this indicates that also the acting mechanisms are different, we decided to proceed as in the previous analyses and as for the holidays in the present one. Hence, we only subtract the largest eigenvalue of the standard correlation matrix. Anticipating the later discussion, we mention that the strongly correlated groups related to the second largest eigenvalue for workdays, i.e. to the largest eigenvalue of the reduced-rank correlation matrix, comprise a large part of the system, but not the system as a whole. This justifies applying the same procedure of analysis as previously, in particular for workdays and holidays.

By removing the largest eigenvalues from the standard correlation matrices, the collective behavior among sections is filtered out. This way reduces the strength of correlations among sections but makes the structures of correlation matrices more distinct, as shown in figure 2. However, the structures are not strongly developed in matrices $\tilde{C}$. To identify correlated groups of sections, we apply the clustering method in section 3.4 to the reduced-rank correlation matrices $\tilde{C}$. This method combines the $k$-means clustering with the significant spectral information. One important step is to determine the number $k$ of clusters. If $k$ is too small, the group information remains hidden, if $k$ is too large, it is blurred. The spectral density (25) of $\tilde{C}$, displayed in figure 3, gives us a basic idea on how the large eigenvalues that deviate from the bulk between the minimal and maximal Marchenko-Pastur eigenvalues are distributed. It is worth mentioning that the eigenvalues of each reduced-rank correlation matrix $\tilde{C}$ from the second smallest to the largest have a one-to-one correspondence with the eigenvalues of its standard correlation matrix $C$ from the smallest to the second largest, but their values are a bit shifted due to the normalization of $\tilde{C}$. For a figure of the spectral density of $C$, we refer the reader to reference [15]. The first three and the first four largest eigenvalues of the reduced-rank correlation matrices for workdays and for holidays, respectively, are much larger than $\tilde{\lambda}_{+}$. As the third and fourth eigenvalues for the case of holidays are close to each other, we use the first three eigenvalues for both cases such that $k - 1 = 3$, i.e., $k = 4$.

We carry out the clustering and find five well-classified section groups, which can be validated by the silhouette values in figure 4. The average silhouette values for both cases are close to 0.5, suggesting appropriate classifications for all sections. To visualize the correlation strength of each group, we reorder all rows and columns of the reduced-rank correlation matrices according to the indices of groups. The order of rows and columns are always the same so that the diagonal elements represent the self-correlation and are equal to one. As a result, the sections within the same group are put together and organized in a descending order of correlations from top to bottom and from left to right in each group. The diagonal blocks of the reordered matrices in figure 2 reveal the internal correlations of each group. We can find at least three strongly correlated groups, for instance, the first four groups for workdays and the middle three groups for holidays. In particular, groups 3 and 4 are obviously anti-correlated for workdays.

4.2 Spectral features of strongly correlated groups

Since the clustering is based on the principal spectral information, we wish to explore how the three largest eigenvalues contribute to each group. Figure 5 displays three-dimensional scatter plots of
Figure 2: The reduced-rank correlation matrices $\tilde{C}$ for workdays (a) and holidays (b), and the reduced-rank correlation matrices $\tilde{C}$ with reordered rows and columns for workdays (c) and holidays (d), where the color indicates the value of correlations and the black lines distinguish group 1 (G1), group 2 (G2), group 3 (G3), group 4 (G4) and group 5 (G5), respectively.

Figure 3: The distributions of spectral density for the reduced-rank correlation matrices $\tilde{C}$, where $\tilde{\lambda}_+$ and $\tilde{\lambda}_-$ are the maximal and minimal Marchenko-Pastur eigenvalues, respectively. $G_i$ stands for group $i$ with $i = 1, 2, 3, 4$ and 5, respectively, and is marked near the corresponding dominant eigenvalue.
Figure 4: The silhouette values for validating the consistency within five clusters. The dot line indicates the averaged silhouette value for workdays (a) and for holidays (b), respectively.

Figure 5: The three-dimensional scatter plots of the eigenvector components corresponding to the largest three eigenvalues of the reduced-rank correlation matrices for workdays (a) and holidays (b), respectively. For a better angle of view, the axes are rotated so that some of axes may not be obvious.
Figure 6: The transposed eigenvector matrices (24), i.e., $\tilde{u}_i^{(k-1)}$, for the largest three eigenvalues, where the eigenvector components along the horizontal axis are ordered corresponding to the reduced-rank correlation matrices in figure 2 (c) and (d). The black lines distinguish group 1 (G1), group 2 (G2), group 3 (G3), group 4 (G4) and group 5 (G5), respectively. The color indicates the value of eigenvector components, i.e., $\tilde{u}_i^{N}$, $\tilde{u}_i^{N-1}$ and $\tilde{u}_i^{N-2}$ in eigenvectors $\tilde{u}_N$, $\tilde{u}_{N-1}$ and $\tilde{u}_{N-2}$, respectively.

Each group in its three-dimensional eigenvector space is located mainly along one or two eigenvectors. We also visualize the eigenvector components of the largest three eigenvalues in figure 6. For each group, at least one eigenvector is strongly occupied by either the positive or the negative components. We then extract the eigenvector matrix (24) of the largest three eigenvalues for each group and rebuild the index for section $i$ from 1 to the total number $q$ of sections in each group, listed in table 2. With regard to the absolute eigenvector components, we can quantify the relative importance of the three eigenvalues by

$$\gamma_j^{(3)} = \frac{1}{q} \sum_{i=1}^{q} \left( |\tilde{u}_i^{(Gg)}|_{N-j+1} - \frac{1}{3} \sum_{j=1}^{3} |\tilde{u}_i^{(Gg)}|_{N-j+1} \right)$$

with $j = 1, 2$ and 3 for the first, the second and the third largest eigenvalues, respectively. The superscript $(Gg)$ stands for group $g$ with $g = 1, 2, 3, 4$ and 5, respectively. A positive value of $\gamma_j^{(3)}$ means that the effect of the $j$-th largest eigenvalue on a group is more than the average effect of the three largest eigenvalues. As a result, the $j$-th largest eigenvalue is dominant in this group.

According to $\gamma_j^{(3)}$, we identify the dominant eigenvalues for each group, as listed in table 2 and marked in figure 3. For the case of workdays, groups 1 and 2 are mainly due to the effect of the

| Group | $q$ | $\langle \tilde{C}_{ij} \rangle_{ij}$ | $\gamma_1^{(3)}$ | $\gamma_2^{(3)}$ | $\gamma_3^{(3)}$ | $q$ | $\langle \tilde{C}_{ij} \rangle_{ij}$ | $\gamma_1^{(3)}$ | $\gamma_2^{(3)}$ | $\gamma_3^{(3)}$ |
|-------|-----|-------------------------------------|----------------|----------------|----------------|-----|-------------------------------------|----------------|----------------|----------------|
| 1     | 361 | 0.1789                             | 0.0136         | -0.0101        | -0.0035        | 282 | 0.036                               | 0.0097         | -0.0034        | -0.0063        |
| 2     | 437 | 0.1048                             | 0.0086         | -0.0084        | -0.0003        | 165 | 0.0928                              | 0.0039         | 0.0061         | -0.0100        |
| 3     | 144 | 0.2062                             | -0.0113        | 0.0256         | -0.0142        | 210 | 0.0721                              | -0.0006        | 0.0139         | -0.0133        |
| 4     | 101 | 0.1784                             | -0.0179        | 0.0241         | -0.0063        | 74  | 0.1088                              | -0.0145        | -0.0236        | 0.0380         |
| 5     | 136 | 0.0639                             | -0.0175        | -0.0115        | 0.0290         | 448 | 0.0262                              | 0.0008         | -0.0037        | 0.0030         |
largest eigenvalue, while group 5 is mostly due to the effect of the third largest eigenvalue. Interestingly, groups 3 and 4 are anti-correlated with each other, as shown in figure 2, and very likely result from the opposite effects of the second largest eigenvalue, revealed by figure 6. We notice that the strongly correlated groups 1 and 2 related to the largest eigenvalue of the reduced-rank correlation matrix comprise, in the case of workdays, a large part of the system, but not the system as a whole. This is a justification, as mentioned before, for our construction of the reduced-rank correlation matrix by subtracting the largest eigenvalue (of the standard correlation matrix) only, although the largest two eigenvalues (of the standard correlation matrix) have similar numerical values in the case of workdays. For the case of holidays, groups 1 and 3 are induced by the effects of the first and the second largest eigenvalues, respectively. In between is group 2 which is dominated by the largest two eigenvalues. Regardless of the remarkable difference in the importance $\gamma_j^{(3)}$, groups 4 and 5 are strongly influenced by the third largest eigenvalues. The aforementioned findings are also supported by the distributions of eigenvector components corresponding to the dominant eigenvalues, as shown in figure 7. The distributions predominantly located on either the positive or the negative side indicate that almost all sections in each group are driven by the same effect represented by the dominant eigenvalue. In particular, the second largest eigenvalue dominates both groups 3 and 4 for workdays, but the corresponding eigenvector components of the two groups are located on opposite sides around zero. This suggests that the opposite effects from the second largest eigenvalue work on the two groups. In addition, group 2 for holidays is dominated by the largest two eigenvalues.

4.3 Geographic features of strongly correlated groups

We project the sections of each group onto the geographic map of NRW, in figure 8 for workdays and in figure 9 for holidays. To better understand the group features, we visualize the data matrices averaged over all workdays and over all holidays in figure 10, where each row shows a time series of velocities $v_n(t)$ for section $n$.

For workdays, the first four groups are strongly correlated. In figure 8, the sections in groups 1 and 2 almost spread over the whole motorway network, where most sections have very high velocities ($v_n(t) > 80$ km/h) during a whole day, i.e., between 0:00 and 23:59, as shown in figure 10. In contrast, the sections in groups 3 and 4 are concentrated in the Rhine-Ruhr metropolitan region with a high population density. The load capacity of each motorway in the Rhine-Ruhr metropolitan region is higher than that in other regions of NRW with a low population density, especially during rush hours. Along the same motorway, e.g., motorway A3 or A57, in figure 8, most sections in group 3 are in opposite directions of the sections in group 4. The difference between the two groups can be traced back to the traffic phases during rush hours. Figure 10 manifests that most sections in group 3 are congested ($v_n(t) < 60$ km/h) during morning rush hours but free ($v_n(t) > 60$ km/h) during...
Identifying strongly correlated groups of sections in a large motorway network

afternoon rush hours. The situation for group 4 is on the contrary. Since the commuter traffic flow dominates in rush hours, a majority of commuters go to work passing through the sections in group 3 during morning rush hours and go back home passing through the sections in group 4 during afternoon rush hours. Taking the section directions into account, we can roughly locate the cities where most commuters work. Two of those cities are Düsseldorf and Cologne. The sections in group 5 are weakly correlated and are scattered over and around the Rhine-Ruhr metropolitan region, see figure 8. They are slightly congested during day time, i.e., between 6:00 and 18:00.

Groups for holidays present regional features in figure 9 – middle region for group 1, center region for group 2, right region for group 3, bottom-left region for group 4 and left region for group 5. In particular, groups 2, 3 and 4 belong to strongly correlated groups. In group 2, most sections are concentrated on motorways A40 and A42, where A40 is the most congested motorway in Germany. The velocities on these sections (80 km/h < \( v_n(t) < 100 \) km/h) are lower than most of those in the other four groups. In contrast, most sections in groups 3 and 4 are far away from the Rhine-Ruhr metropolitan region and their velocities (\( v_n(t) > 120 \) km/h) are remarkably high during day time.

4.4 Dominant eigenvalues related to geographic distributions of sections

The spectral features clarify the contributions of the dominant eigenvalues to each group, while the geographic distributions reveal the section-concentrated regions for each group. We now associate the dominant eigenvalues with the geographic distributions of motorway sections for workdays. They are, compared with the case of the holidays, more interesting, diverse and functional. A merger using dominant eigenvalues can be carried out.

As shown in figure 11, by combining groups 1 and 2, the sections for the largest eigenvalue \( \tilde{\lambda}_N \) are distributed on the whole state and most of them are in a free traffic phase at any time of the day. Hence, the largest eigenvalue \( \tilde{\lambda}_N \) is related to the free traffic phase during a whole day. Regardless of the anti-correlation between groups 3 and 4, their sections for the second largest eigenvalue \( \tilde{\lambda}_{N-1} \) are distributed in the Rhine-Ruhr metropolitan region with a high population density. As discussed above, the sections in the two groups are congested during morning or afternoon rush hours due to the commuter traffic flow. Hence, the second largest eigenvalue \( \tilde{\lambda}_{N-1} \) is related to the congested traffic phases during rush hours. The sections for the third largest eigenvalue are the ones in group 5 and share the same geographic features with group 5. Hence, the third largest eigenvalue is related to the slightly congested traffic phase during day time.

5 Conclusions

To identify strongly correlated groups of sections in the NRW motorway network, we developed and applied a clustering method using spectral information. The key idea is to go to the moving frame of the largest eigenvalue and to construct a reduced-rank correlation matrix which reveals the strongly correlated groups in a much clearer fashion. Furthermore, clustering based on the spectral information was implemented by an optimized \( k \)-means method.

We applied the spectral approach above-mentioned to empirical data from 1179 motorway sections in NRW, Germany, and classified all sections into five groups. According to the correlation strength, the first four groups for the case of workdays and the middle three groups for the case of holidays were identified as the strongly correlated groups. For the case of workdays, the first two groups are dominated by the largest eigenvalue of the reduced-rank correlation matrix. Their sections spread to the whole state and most of sections are under a free traffic phase with very high velocities during a whole day. The third and the fourth groups are dominated by the second largest eigenvalues. Their sections are concentrated in the Rhine-Ruhr metropolitan region of NRW. Most sections in the third (fourth) group are congested (free) during morning rush hours but free (congested) during afternoon rush hours. The fifth group is a weakly correlated group and dominated by the third largest eigenvalue. Its sections are scattered over and around the Rhine-Ruhr metropolitan region and are slightly congested during day time. For the case of holidays, the groups can be separated by regions. All groups correspond to high velocities except the second one whose sections are
Figure 8: Five groups of motorway sections of NRW in Germany for workdays. The green background indicates the population density of districts in NRW. For the information and the data source of the base map, see figure 1.
Identifying strongly correlated groups of sections in a large motorway network

Figure 9: Five groups of motorway sections of NRW in Germany for holidays. The green background indicates the population density of districts in NRW. For the information and the data source of the base map, see figure 1.
Identifying strongly correlated groups of sections in a large motorway network

Figure 10: Data matrices of dimension 1179 $\times$ 1440. Each row shows a time series of velocities $v_n(t)$ for section $n$. The data matrices are averaged over all workdays (a) and over all holidays (b), respectively. The rows of data matrices (a) and (b) are ordered corresponding to the reduced-rank correlation matrices in figure 2 (c) and (d), respectively. The black lines distinguish group 1 (G1), group 2 (G2), group 3 (G3), group 4 (G4) and group 5 (G5), respectively. The color indicates the value of $v_n(t)$ in units of km/h.

mainly concentrated on the motorways A40 and A42. The congestion is almost absent on the sections in all groups for the holidays.

The approach developed in this study led to a remarkably clear identification and separation of the strongly correlated groups of motorway sections. In particular, the third and the fourth groups identified for workdays are strongly related to the commuter traffic flows during rush hours. In contrast to others, the sections in these two groups are more likely to be critical bottlenecks with respect to the load of motorways. To improve the traffic efficiency, it is better to bypass these sections when determining alternative routes during rush hours. These sections also should be given a priority if any road improvement would be implemented to relieve the load of motorways and enhance the traffic efficiency. From a more general viewpoint, our study of the correlation structure provides completely new information on the network and the dynamics on it. Our approach can also be applied to other correlated and non-stationary complex systems for identifying strongly correlated groups.

Acknowledgements

We gratefully acknowledge funding via the grant “Korrelationen und deren Dynamik in Autobahnnetzen”, Deutsche Forschungs gemeinschaft (DFG, 418382724). We thank Strassen.NRW for providing the empirical traffic data. We also thank Sebastian Gartzke, Anton Josef Heckens, Daniel Waltner and Henrik Bette for fruitful discussions.

Author contributions

T.G. and M.S. proposed the research. S.W. and T.G. developed the methods of analysis. S.W. performed all the calculations. S.W. and T.G. wrote the manuscript with the input from M.S. All authors contributed equally to analyzing the results and reviewing the paper.

References

[1] Robert M May. Will a large complex system be stable? Nature, 238(5364):413–414, 1972.
Identifying strongly correlated groups of sections in a large motorway network

Figure 11: The geographic distributions of motorway sections of NRW in Germany with the dominant eigenvalues for workdays. The green background indicates the population density of districts in NRW. For the information and the data source of the base map, see figure 1.

[2] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? Eur. J. Philos. Sci., 3(1):33–67, 2013.

[3] Shanshan Wang, Sebastian Gartzke, Michael Schreckenberg, and Thomas Guhr. Quasi-stationary states in temporal correlations for traffic systems: Cologne orbital motorway as an example. J. Stat. Mech. Theory Exp., 2020(10):103404, 2020.

[4] Ross Baldick, Badrul Chowdhury, Ian Dobson, Zhaoyang Dong, et al. Initial review of methods for cascading failure analysis in electric power transmission systems ieee pes cam task force on understanding, prediction, mitigation and restoration of cascading failures. In 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pages 1–8. IEEE, 2008.

[5] Zoltán Toroczkai and Kevin E Bassler. Jamming is limited in scale-free systems. Nature, 428(6984):716–716, 2004.

[6] Li Daqing, Jiang Yinan, Kang Rui, and Shlomo Havlin. Spatial correlation analysis of cascading failures: congestions and blackouts. Sci. Rep., 4(1):1–6, 2014.

[7] Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Andrea Gabrielli, and Michelangelo Puliga. Reconstructing a credit network. Nat. Phys., 9(3):125–126, 2013.
Identifying strongly correlated groups of sections in a large motorway network

Daqing Li, Bowen Fu, Yunpeng Wang, Guangquan Lu, Yehiel Berezin, H Eugene Stanley, and Shlomo Havlin. Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proc. Natl. Acad. Sci. U.S.A.*, 112(3):669–672, 2015.

Guanwen Zeng, Daqing Li, Shengmin Guo, Liang Gao, Ziyou Gao, H Eugene Stanley, and Shlomo Havlin. Switch between critical percolation modes in city traffic dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 116(1):23–28, 2019.

Guanwen Zeng, Jianxi Gao, Louis Shekhtman, Shengmin Guo, Weifeng Lv, Jianjun Wu, Hao Liu, Orr Levy, Daqing Li, Ziyou Gao, et al. Multiple metastable network states in urban traffic. *Proc. Natl. Acad. Sci. U.S.A.*, 117(30):17528–17534, 2020.

Yanhui Liu, Parameswaran Gopikrishnan, Cizeau, Meyer, Peng, and H Eugene Stanley. Statistical properties of the volatility of price fluctuations. *Phys. Rev. E*, 60(2):1390, 1999.

Parameswaran Gopikrishnan, Bernd Rosenow, Vasiliki Plourou, and H Eugene Stanley. Quantifying and interpreting collective behavior in financial markets. *Phys. Rev. E*, 64(3):035106, 2001.

Vasiliki Plourou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65(6):066126, 2002.

Anton J Heckens, Sebastian M Krause, and Thomas Guhr. Uncovering the dynamics of correlation structures relative to the collective market motion. *J. Stat. Mech. Theory Exp.*, 2020(10):103402, 2020.

Shanshan Wang, Sebastian Gartzke, Michael Schreckenberg, and Thomas Guhr. Collective behavior in the North Rhine-Westphalia motorway network. *J. Stat. Mech. Theory Exp.*, 2021(12):123401, 2021.

Bundesamt für Kartographie und Geodäsie. Verwaltungsgebiete 1:2 500 000, Stand 01.01. (VG2500). https://gdz.bkg.bund.de/index.php/default/verwaltungsgebiete-1-2-500-000-stand-01-01-vg2500.html, 2020.

Das Datenportal für Deutschland. Data licence Germany – attribution – version 2.0. http://www.govdata.de/dl-de/by-2-0, 2021.

Statistische Ämter des Bundes und der Länder, Deutschland. Regionalatlas Deutschland. https://regionalatlas.statistikportal.de, 2021.

OpenStreetMap. Copyright and License. https://www.openstreetmap.org/copyright, 2021.

Open Knowledge Foundation. Open Data Commons Open Database License (ODbL) v1.0. https://opendatacommons.org/licenses/odbl/1-0/, 2021.

QGIS. Documentation for QGIS 3.4. https://docs.qgis.org/3.4/en/docs/, 2020.

Hirdesh K Pharasi, Kiran Sharma, Anirban Chakraborti, and Thomas H Seligman. Complex market dynamics in the light of random matrix theory. In *New Perspectives and Challenges in Econophysics and Sociophysics*, pages 13–34. Springer, 2019.

Yuriy Stepanov, Philip Rinn, Thomas Guhr, Joachim Peinke, and Rudi Schäfer. Stability and hierarchy of quasi-stationary states: financial markets as an example. *J. Stat. Mech. Theory Exp.*, 2015(8):P08011, 2015.

J Scott Goldstein and Irving S Reed. Reduced-rank adaptive filtering. *IEEE Trans. Signal Process.*, 45(2):492–496, 1997.
[25] Anirban Chakraborti, Kiran Sharma, Hirdesh K Pharasi, K Shuvo Bakar, Sourish Das, and Thomas H Seligman. Emerging spectra characterization of catastrophic instabilities in complex systems. *New J. Phys.*, 22(6):063043, 2020.

[26] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 29, 2004.

[27] Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[28] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[29] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

[30] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *Int. J. Theor. Appl. Finance*, 3(03):391–397, 2000.

[31] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, LA Nunes Amaral, and H Eugene Stanley. A random matrix theory approach to financial cross-correlations. *Physica A*, 287(3-4):374–382, 2000.

[32] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, volume 344. John Wiley & Sons, 2009.

[33] Boris S Kerner. *The physics of traffic: empirical freeway pattern features, engineering applications, and theory*. Springer, 2012.