Data adaptive flow directed PCA : Application in Geum River data

Kyusoon Kim

Seoul National University

kyu9510@snu.ac.kr

March 17, 2021

Overview

1. Introduction
2. Background
3. Methodology
4. Real Data Analysis
5. Conclusion

Introduction

- River water quality monitoring is necessary because it reflects environmental status which is closely related to human society.
- Identifying spatial and temporal changes in the river quality plays a major role of managing land and water resources and preventing environmental pollution.
- The goal of this study is to figure out spatio-temporal patterns of river network data by reducing dimensionality with PCA.
- How can we apply PCA on flow-directed river network data?
- In this study, we apply the flow-directed PCA suggested by Gallacher et al.(2017) to Geum River data and develop the suggested method.

Principal Component Analysis

- Principal Component Analysis (PCA) is a linear dimension reduction technique that gives a set of direction vectors of maximal variances.
- PCA finds uncorrelated new variables, the principal components (PCs), which are linear functions of those in the original dataset.
- PCA can be used to identify patterns in the data.

River Network data

- In water quality monitoring case, observed values are located on the river network.

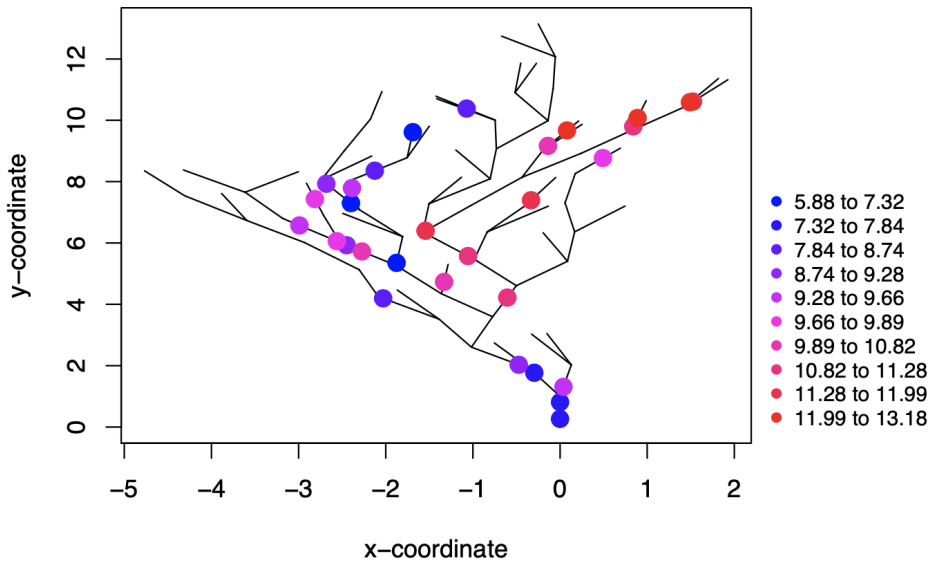


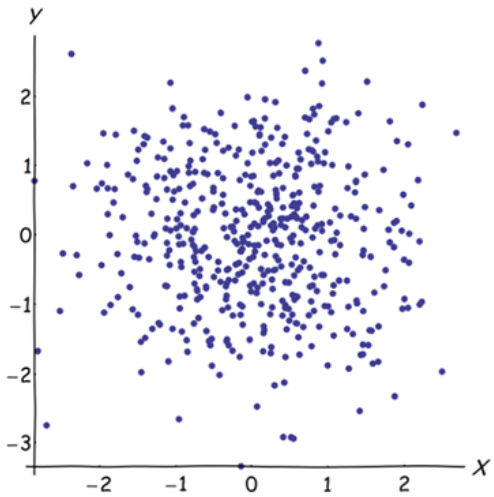
Figure: Example of river network data (Gallacher et al., 2017)

River Network data

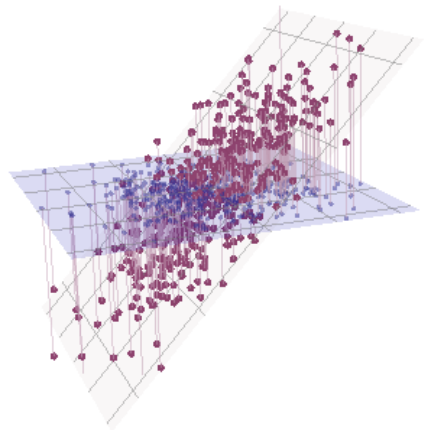
- There are several distinct characteristics of the river network.
  - River network has a direction of flow and is connected by river flow.
  - Observed values between sites at subsequent time points will be related.
- These characteristics cause spatial and temporal correlation over the river network.

Necessity of removing highly correlated variables before PCA

- If we add a third variable close to the second variable ?



(a) Data in 2-dimension



(b) Data in 3-dimension

Figure: Influence of highly correlated variables

- Initially, the contributions of two variables are almost equal
- After adding one variable, the variance explained by the first PC is twice the amount of the other.

Necessity of removing highly correlated variables before PCA

- If we continue to add correlated variables, only one principal component would be worth considering.
- Therefore, highly correlated variables can cause the PCA to overemphasize their contribution and mask identification of important pattern.
- We will use a weighted PCA on the river network in order to adjust PCA for network structure and temporal autocorrelation.

Weighted PCA

- For column mean centered $n \times p$ matrix X, a $p \times p$ column weight matrix Ω and $n \times n$ row weight matrix Φ are constructed and then PCA is applied to $\tilde{X} = \Phi X \Omega = \tilde{U} \tilde{D} \tilde{V}^T$.
- The PCs and loadings related to X can be obtained by backtransformation.
- Consequently, the PCs are $X \Omega \tilde{V}$ and the loadings are $\Omega^{-1\,T} \tilde{V}$

Spatial and Temporal weights

- Shreve stream ordering is one of the methods of assigning a numeric order to link in a stream network.

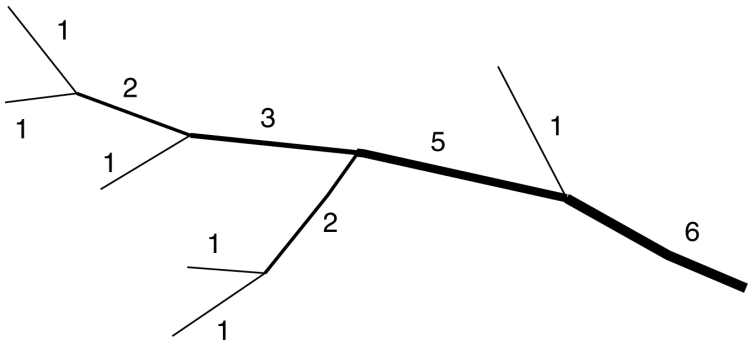


Figure: Shreve stream order

- Additive Function Value (AFV) of monitoring site M is defined as

$$\frac{\textit{Shreve order of the stream segment containing } M}{\textit{Shreve order of the river mouth}}$$

Spatial and Temporal weights

- Suppose data matrix X is $n \times p$ matrix with n time points and p monitoring sites.
- A $p \times p$ asymmetric matrix of spatial weights S can be constructed by calculating $\pi_{u,d} = \sqrt{\frac{AFV_u}{AFV_d}}$.

$$S_{d,u} = \begin{cases} \pi_{u,d}, & \text{if u and d are flow connected and represent upstream site and downstream site respectively} \\ 0, & \text{if u and d are not flow connected or u and d are flow connected but represent downstream site and upstream site respectively} \end{cases}$$

Spatial and Temporal weights

- AR(1) model is sufficient to capture autocorrelation in water quality time series (Clement et al. (2006) and Andrés Houseman (2005)).
- A $n \times n$ symmetric matrix of temporal weights T is defined as

$$T_{ij} = \rho^{|i-j|}$$

  where ρ is the strength of correlation between observed values at consecutive time points.
- $S^{-1/2}$ and $T^{-1/2}$ is used to adjust PCA for spatial and temporal autorcorrelation respectively.

Example of river network data

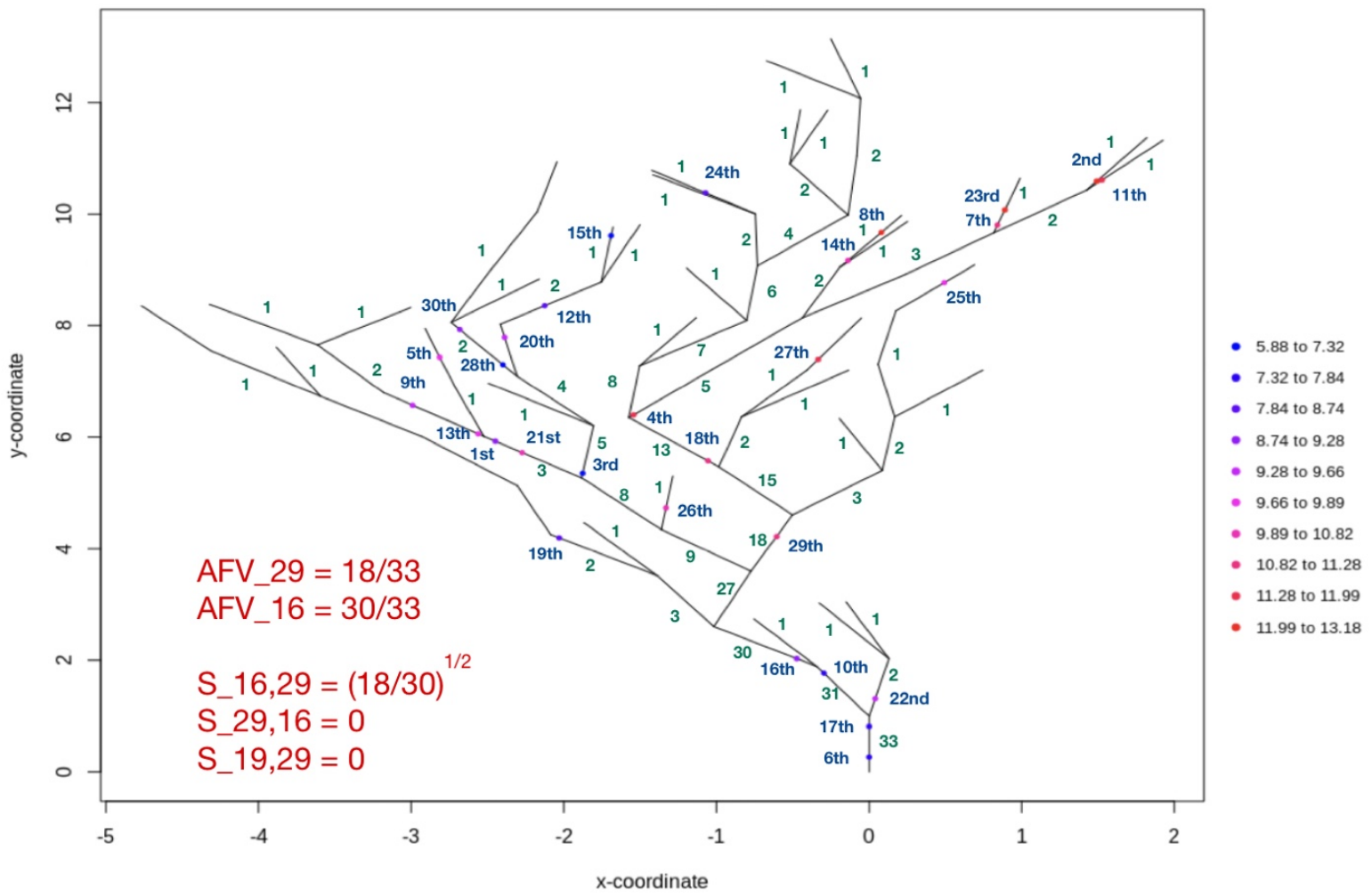


Figure: Example of river network data

S-mode PCA

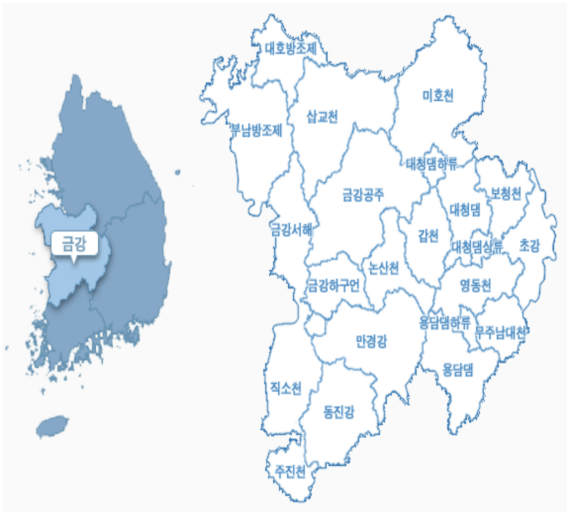
- S-mode PCA aims to estimate dominant temporal patterns and figure out which sites show similar temporal patterns.
- Data matrix X is arranged so that each column (variable) represents a monitoring site and each row (observation) represents an ordered time point.

T-mode PCA

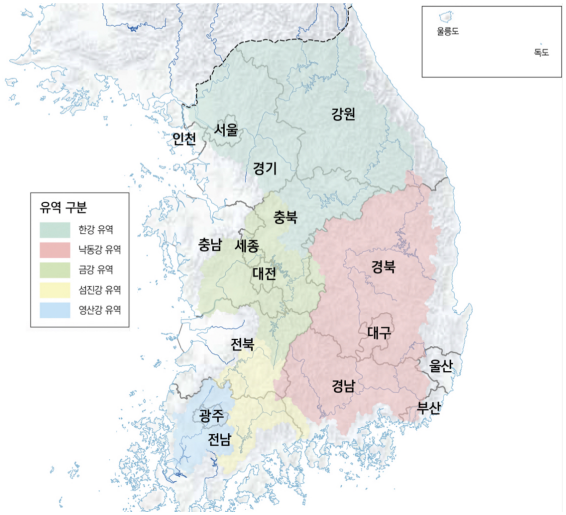
- T-mode PCA aims to identify spatial patterns and the associated time points at which these spatial patterns occur.
- Data matrix X is arranged so that each column (variable) represents an ordered time point and each row (observation) represents a monitoring site.
- The presence of more than one dominant spatial pattern suggests a change in the spatial pattern over time.
- A single dominant spatial pattern means that any specific pattern has remained stable over time.

Geum river Data

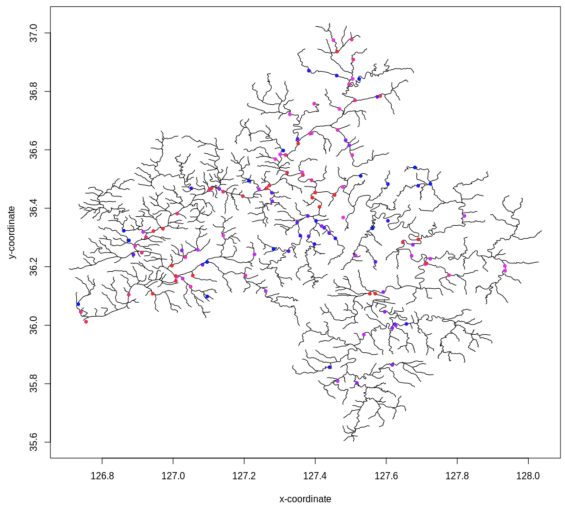
- We will use Total Organic Carbon (TOC) data, the amount of carbon found in an organic compound, which is often used as an indicator of water quality.



(a) Geum river basin



(b) Geum catchment area



(c) Geum river TOC data

Figure: Geum river data

Geum river Data

- Daily TOC data were measured at 127 monitoring sites from the winter of 2011 till the fall of 2017.
- A natural log transformation of TOC observations was taken to stabilize the variance over time and across the network.
- Data measured in summer are of specific interest because the average of TOC concentration is typically higher in summer.

Missing value imputation

- River network often has a lot of missing values due to several reasons.
  - Water monitoring strongly depends on automatic acquisition systems.
  - Automatic sensing devices are easily exposed to harsh ecological conditions so that they are subject to physical destruction and battery drainage.
- Missing value imputation can be achieved by an iterative PCA algorithm.

Application : T-mode PCA & S-mode PCA

- For T-mode PCA, we will use annual average summer data.
- However, for S-mode PCA, rows represents time points so that there are more variables than observations which result in the poor result when doing PCA.
- Therefore, for S-mode PCA, we will use monthly average data.

Result : T-mode PCA

- The percentage of the variance explained by the first principal component has decreased compared to unweighted T-mode PCA.
- It means adjusting PCA for spatial and temporal correlation has removed some of the correlation.
- Adjusting for temporal correlation has had a greater effect on the results than adjusting for spatial correlation.

PCA	PC1(%)	PC2(%)	PC3(%)	$var_3(\%)$	k	$var_k(\%)$	ϵ_k
$TPCA_{uw}$	91.4	3.1	2.5	97.0	1	91.4	1179.2
$TPCA_S$	88.2	4.7	3.4	96.3	2	92.9	1174.1
$TPCA_T$	80.0	8.9	5.0	93.9	3	93.9	1170.5

Table: T-mode PCA result. k is the number of principal components retained to explain at least 90% of the variance. ϵ_k is a reconstruction error.

Result : T-mode PCA

- The loadings for the first principal component are all of the same sign and of similar magnitude, thus the first PC represents the average spatial pattern over all years.
- The second PC represents a contrast between 2016 and other years.

Year	UW_1	UW_2	SW_1	SW_2	TW_1	TW_2
2012	0.33	-0.32	0.34	-0.34	0.46	-0.15
2013	0.42	-0.19	0.43	-0.23	0.57	**0.18**
2014	0.38	-0.07	0.37	-0.02	0.52	-0.03
2015	0.45	-0.02	0.44	0.03	0.62	-0.14
2016	0.43	**0.86**	0.42	**0.85**	0.58	**0.64**
2017	0.43	-0.34	0.44	-0.34	0.59	-0.27

Table: Loadings for unweighted (UW), spatially weighted (SW) and temporally weighted (TW) T-mode PCA

Result : T-mode PCA

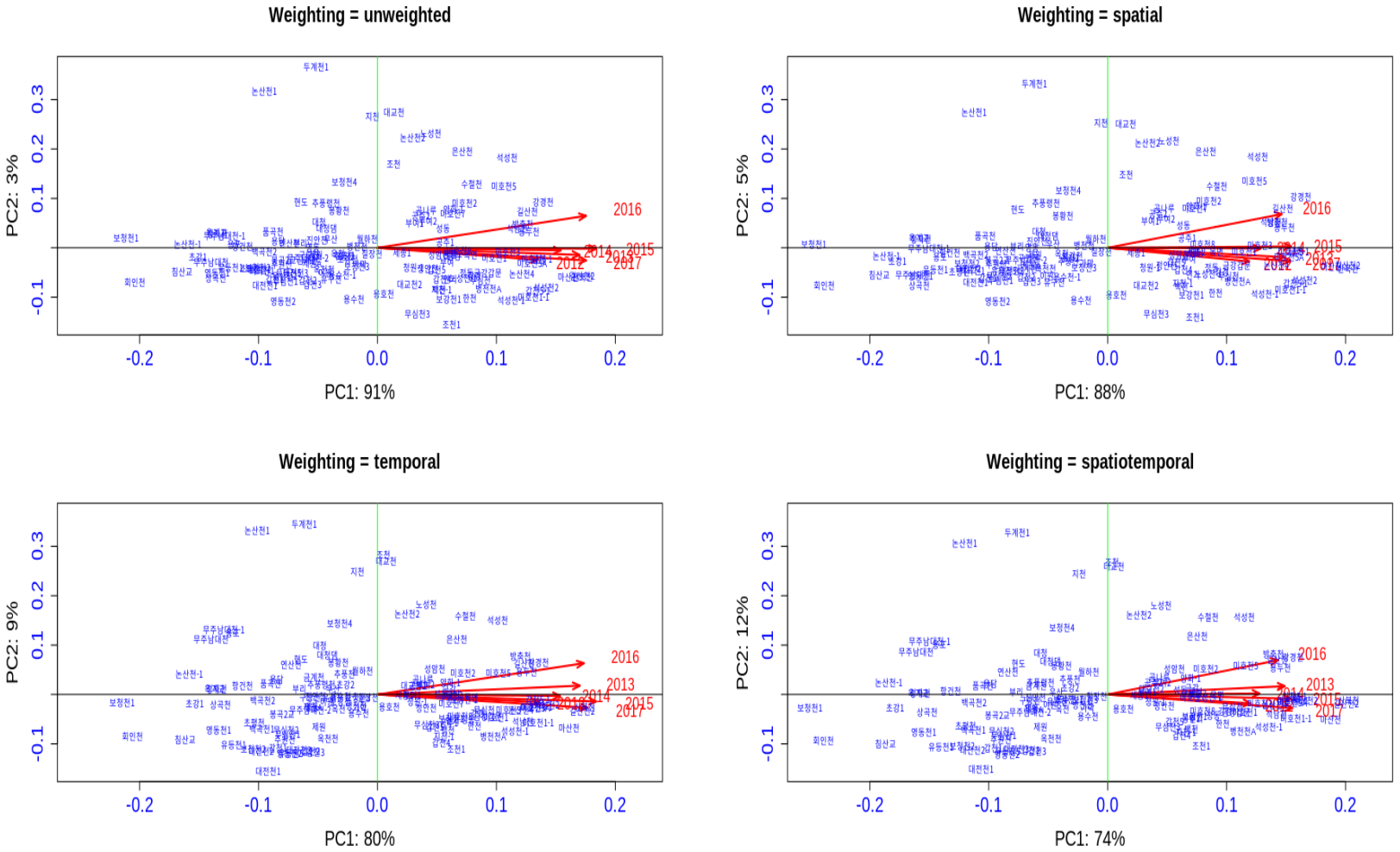


Figure: Biplots for the first two T-mode principal components
Blue points are scores and red arrows are loadings.

Result : T-mode PCA

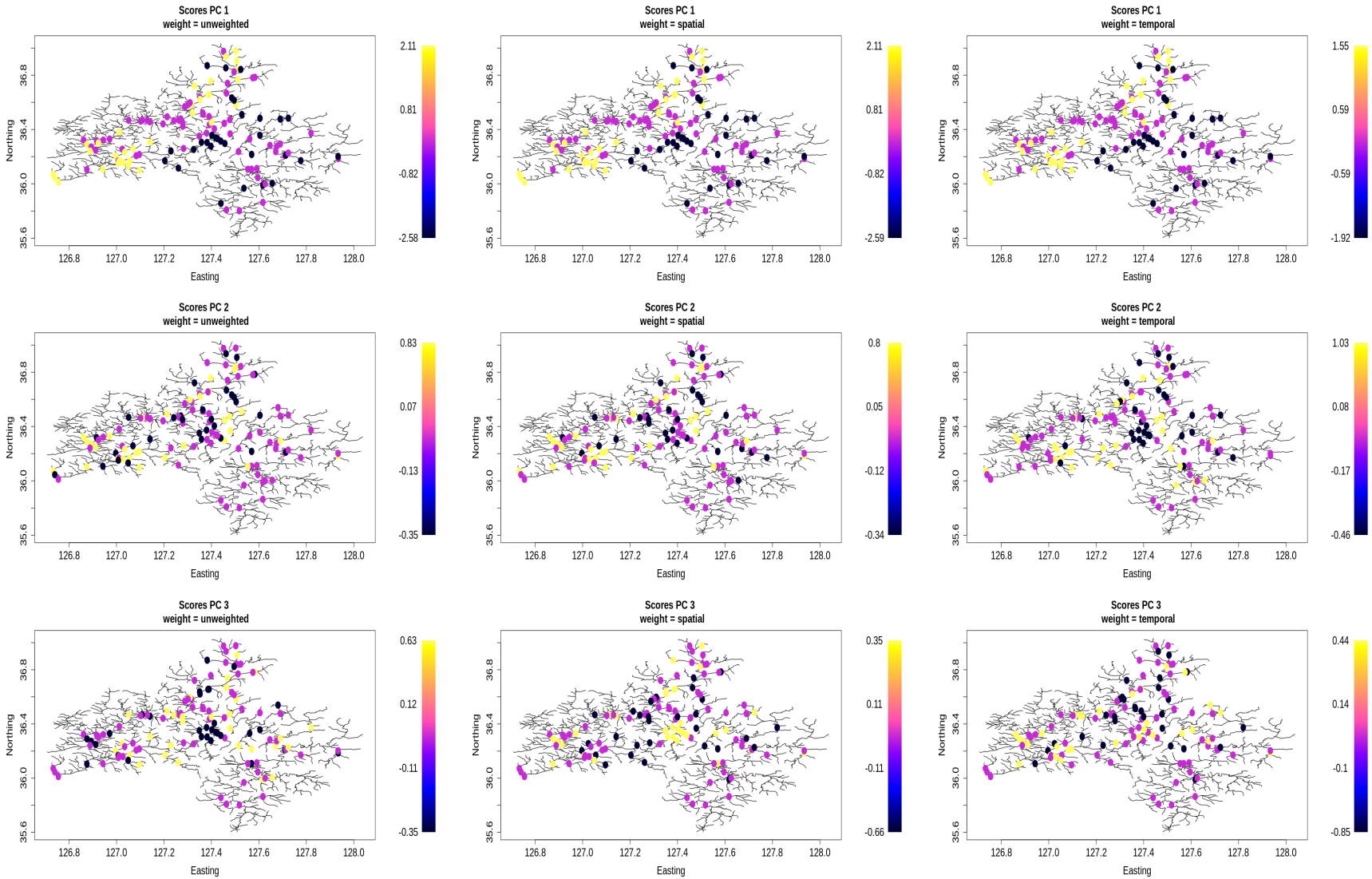
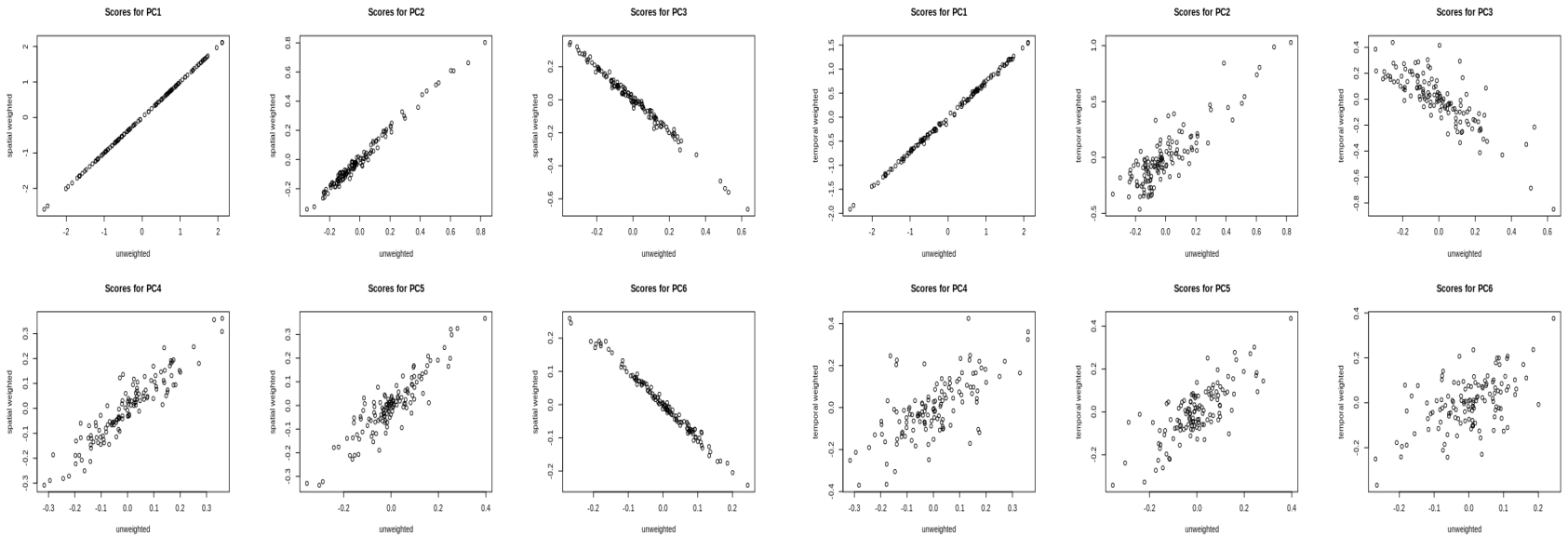


Figure: Loadings for first three S-mode principal components

Result : T-mode PCA

- Sites in the top and bottom right quadrants (North-West of the catchment) have high log(TOC) every summer, while sites in the top and bottom left quadrants (South-East of the catchment) have low(TOC) every summer.
- Difference in scores between unweighted and weighted PCA can be found in the higher degree PCs that explain small proportion of the variance in the data.

Result : T-mode PCA



(a) Unweighted vs Spatial weighted **(b)** Unweighted vs Temporal weighted

Figure: Comparison of Principal component scores

- Spatial and temporal weights reflect correlation in the noise structure after removing trend.

Result : T-mode PCA

- The glyph plots for T-mode PCA show that when temporal correlation is accounted for, the second and third PCs capture spatial patterns that were not visible in the unweighted PCA.

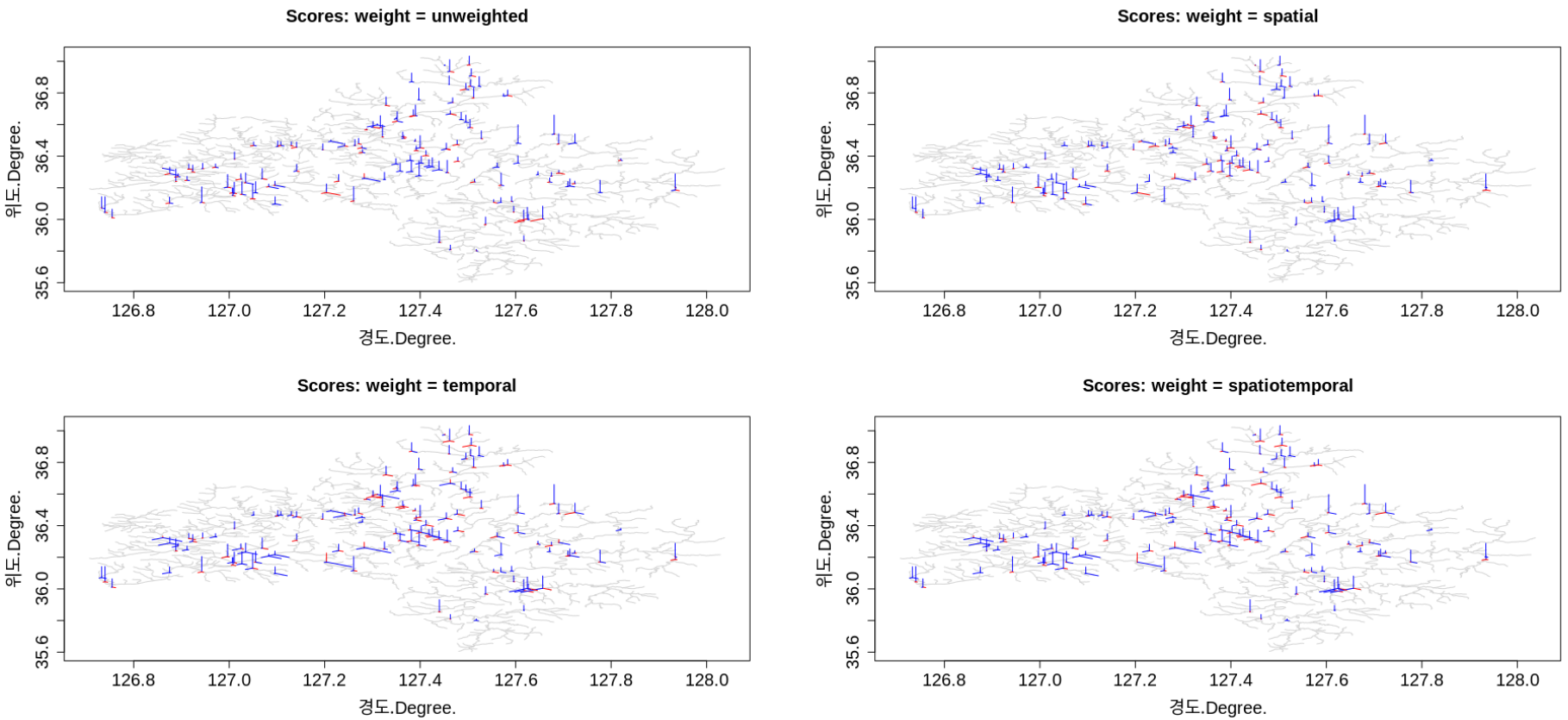


Figure: Glyph plots of T-mode PCA

Summary : T-mode PCA

- A dominant spatial pattern of log(TOC) has remained stable over time.
- Monitoring sites in the North-West of the Geum catchment such as Daejeon and Sejong showed higher level of TOC than in the South-East of the catchment because of the development and high population density.
- The second T-mode principal component explained a small amount of the total variance but represented a contrast between 2016 and other years.
- T-mode PCA has discovered spatial patterns masked by autocorrelation.

Result : S-mode PCA

- The percentage of the variance explained by the first principal component has decreased compared to unweighted T-mode PCA.
- It means adjusting PCA for spatial and temporal correlation has removed some of the correlation.

PCA	PC1(%)	PC2(%)	PC3(%)	$var_3(\%)$	k	$var_k(\%)$	ϵ_k
$SPCA_{uw}$	41.1	14.3	9.3	64.8	4	71.7	11803.5
$SPCA_S$	35.3	15.0	12.5	62.8	5	72.4	11758.6
$SPCA_{ST}$	25.9	14.4	11.7	52.0	8	70.5	11728.7

Table: S-mode PCA result. k is the number of principal components retained to explain at least 70% of the variance. ϵ_k is a reconstruction error.

Result : S-mode PCA

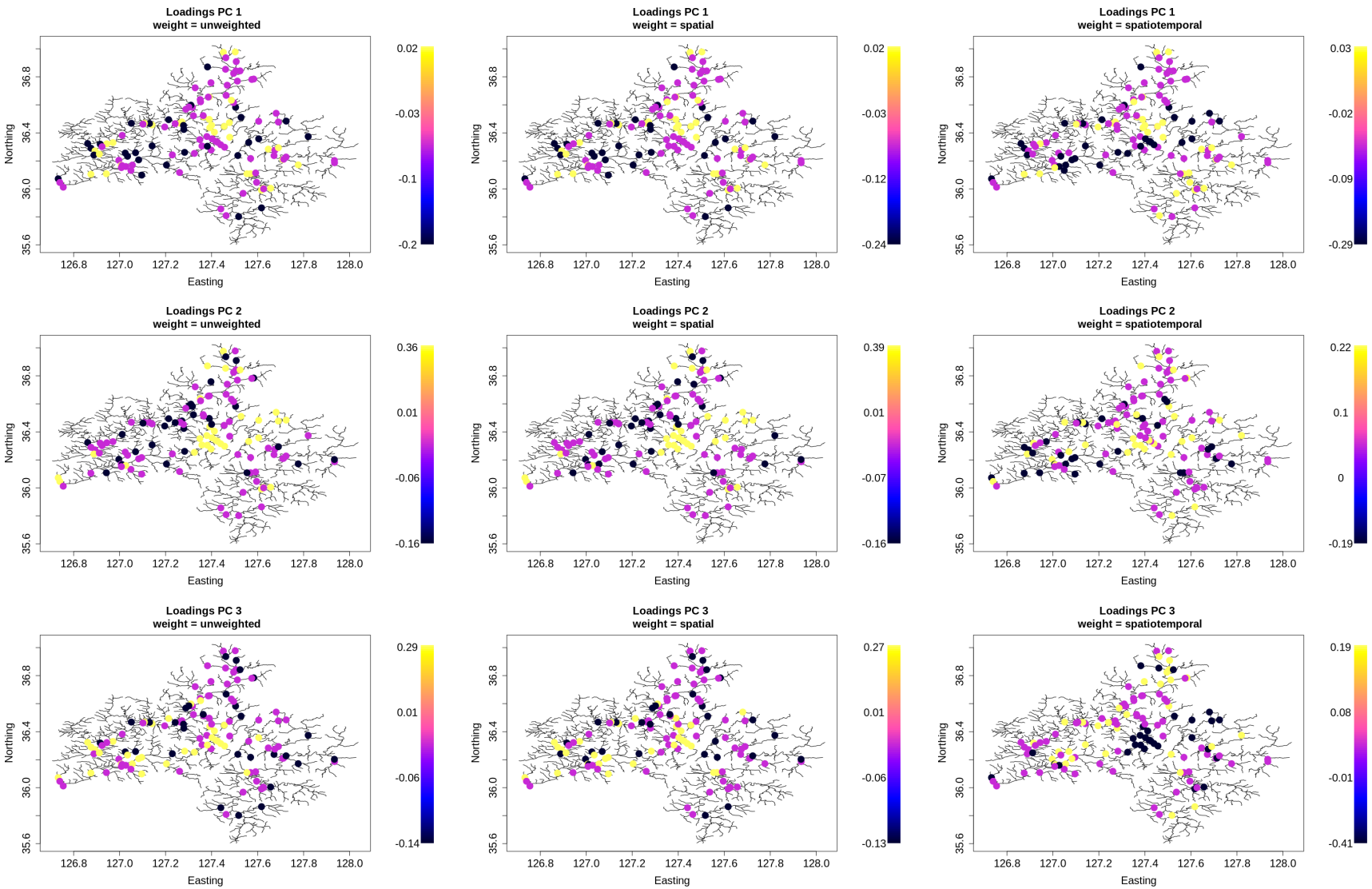


Figure: Loadings for first three S-mode principal components

Result : S-mode PCA

- For S-mode PCA, we can interpret that sites with loadings of the same sign and similar magnitude exhibit similar temporal pattern.
- A high loading means that the temporal pattern described by the PC related to the loading can be found that site.
- The distribution of loadings is very similar for the first principal component between three weighting schemes and the difference can be found for the third principal component after removing trend.

Result : S-mode PCA

- The glyph plots show the distinction between influential and less influential monitoring sites.

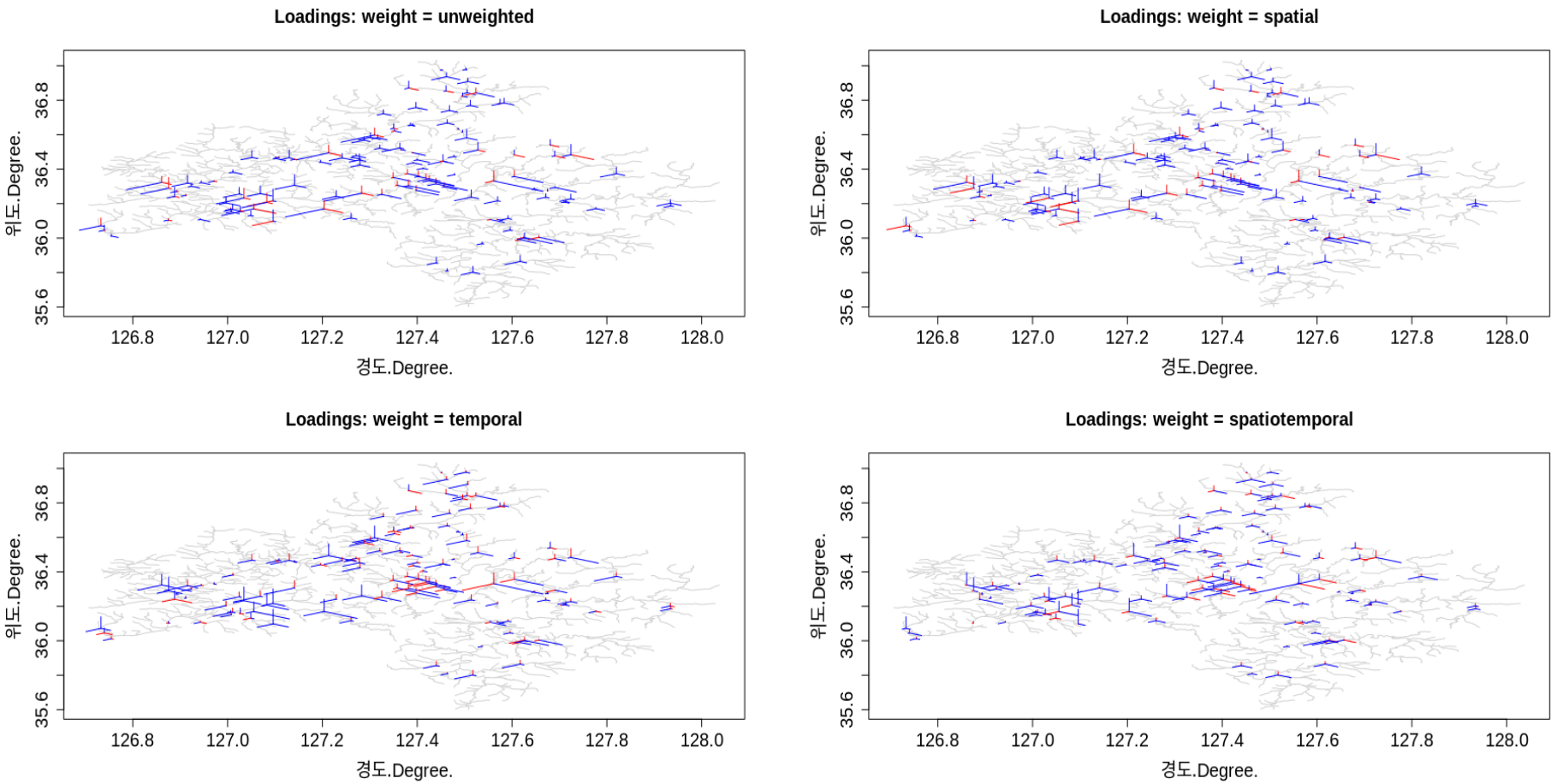


Figure: Glyph plots of S-mode PCA

Summary : S-mode PCA

- Monitoring sites showing similar temporal patterns can be identified through S-mode PCA.
- We can figure out which monitoring sites contribute little to the variance of the river.
- S-mode PCA has discovered temporal patterns masked by autocorrelation.

Remaining questions

- Missing value imputation problem
  - Is an Iteraitive PCA Algorithm appropriate for the flow-directed data?
  - Data adaptive missing value imputation
- Spatial weight construction problem
  - Is Shreve stream order effective to construct spatial weights?
  - Is it okay to give spatial weights in the same way across the whole river network system?
  - Data adaptive spatial weights
- Temporal weight construction problem
  - Is AR(1) structure sufficient to model temporal correlation?
  - Data adaptive temporal weights

Remaining questions

- Missing value imputation problem
  - Is an Iteraitive PCA Algorithm appropriate for the flow-directed data?
  - Data adaptive missing value imputation
- Spatial weight construction problem
  - **Is Shreve stream order effective to construct spatial weights?**
  - Is it okay to give spatial weights in the same way across the whole river network system?
  - Data adaptive spatial weights
- Temporal weight construction problem
  - Is AR(1) structure sufficient to model temporal correlation?
  - Data adaptive temporal weights

Modification of spatial weights construction

- Does same shreve stream order mean same flow volume?
  Larger upstream segment will make a greater impact on downstream segment.

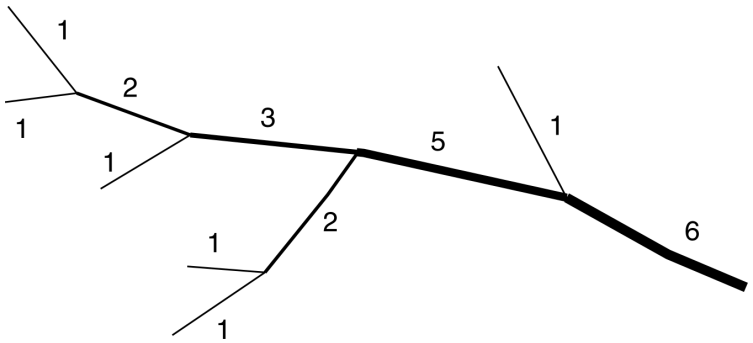


Figure: Shreve stream order

- We can assume that the flow of the upper-most stream segments is proportional to their lengths.
- Flow volume of each segment can be calculated with the same method calculating shreve stream order.

Modification of spatial weights construction

PCA	PC1(%)	PC2(%)	PC3(%)	$var_3(\%)$	k	$var_k(\%)$	ϵ_k
$TPCA_{uw}$	91.4	3.1	2.5	97.0	1	91.4	1179.2
$TPCA_S$	88.2	4.7	3.4	96.3	2	92.9	1174.1
$TPCA'_S$	88.3	4.7	3.4	96.4	2	93.0	1174.1

Table: T-mode PCA result. k is the number of principal components retained to explain at least 90% of the variance. ϵ_k is a reconstruction error.

PCA	PC1(%)	PC2(%)	PC3(%)	$var_3(\%)$	k	$var_k(\%)$	ϵ_k
$SPCA_{uw}$	41.1	14.3	9.3	64.8	4	71.7	11803.5
$SPCA_S$	35.3	15.0	12.5	62.8	5	72.4	11758.6
$SPCA'_S$	35.4	15.0	12.4	62.7	5	72.4	11758.5

Table: S-mode PCA result. k is the number of principal components retained to explain at least 70% of the variance. ϵ_k is a reconstruction error.

Conclusion

- Implementation of PCA on the flow-directed network should be adjusted for spatial and temporal correlation.
- T-mode PCA and S-mode PCA can be used to identify spatial and temporal patterns masked by spatial and temporal correlation respectively.
- Estimating flow volume of the upper-most stream segments proportional to their stream length was not effective.

Future works

- Missing value imputation method based on the flow directed data might be needed to deal with a large number of missing values.
- River network has inhomogeneous covariance structure so that weights reflecting inhomogeneity can be constructed.
- Data adaptive time series modeling instead of assuming AR(1) structure of river network may improve the performance of weighted PCA.

References

Gallacher, K., et al. (2017). Flow-directed PCA for monitoring networks. *Environmetrics*, **28(2)**, e2434.

Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P. (2006). Spatio-temporal statistical models for river monitoring networks. *Water science and technology*, **53(1)**, 9—15.

Andrés Houseman, E. (2005). A robust regression model for a first-order autoregressive time series with unequal spacing: Application to water monitoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54(4)**, 769-780.

Andrés Houseman, E. (2005). A robust regression model for a first-order autoregressive time series with unequal spacing: Application to water monitoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54(4)**, 769-780.

References

Gallacher, K. (2016). Using river network structure to improve estimation of common temporal patterns. *Doctoral dissertation, University of Glasgow*.

Park, S. (2019). Multiscale Analysis of Spatio-Temporal Data. *Doctoral dissertation, Seoul National University*.

Thank you