Creating_demo_data

Kelly Gallacher and Claire Miller and Marian Scott

2016-04-28

1 Introduction

This document explains how the `demoY`, `demoYmiss`, and `demoNet` data sets were created for use in the stpca package.

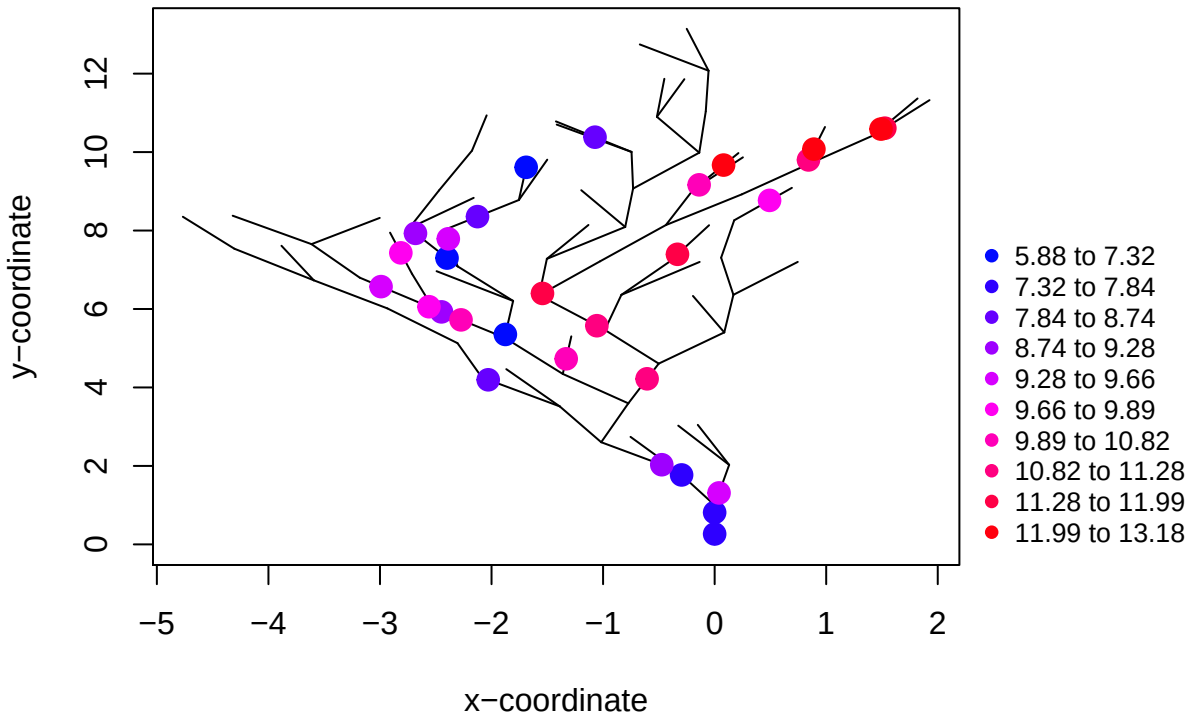
2 SSN object

A river network was simulated as an object of class`"SpatialStreamNetwork"` (SSN object) and was created using functions within the SSN package in R. The SSN object has 80 stream segments, $j = 30$ monitoring sites with simualated 'observed' values, and 30 prediction locations with no observed values. The model used to simulate observed values y_j at a single time point is an intercept only model (no covariates or spatial trend):

$$y_j = \mu + \varepsilon_j$$
$$\varepsilon_j \sim N(0, \Sigma)$$

In this model, $\mu = 10$, Σ is the spatial covariance matrix combining Epanechnikov Tail-up, Gaussian Euclidean, and nugget (iid N(0,1)) variance components. The tail-up component has range = 2, and $\sigma^2_{TU} = 100$. The Euclidean component has range = 2, $\sigma^2_{Euc} = 50$. The additive function for the SSN model is based on Shreve's stream order. These parameters give variance components of $R^2 = 0$, tail-up = 57%, Euc = 28%, and nugget = 15%.

The network, coloured by observed values ("Sim_Values"), is shown below:



3 Variable for stratified sampling

A variable `"strata"` was added to the observed locations, to be used for stratified sampling when investigating the effect of reducing the size of the monitoring network. Each monitoring location was randomly allocated to strata = 1 or strata = 2, and resulted in 12 observed monitoring sites in strata = 1 (18 in strata = 2), and for prediction locations there were 16 in strata = 1 (14 in strata = 2).

4 Variable for weighted sampling

A variable `"weight"` was added to the observed locations, to be used for weighted sampling when investigating the effect of reducing the size of the monitoring network. Each monitoring location was assigned a probability of being included in the sampled network with a random draw from a uniform distribution, with lower boundary = 0.01 and upper boundary = 0.99. This ensures that no single site is guaranteed to be omitted from every sampled network, and no single site is guaranteed to be included in every sampled network.

The simulated river network can be found in `demoNet`.

5 Spatially correlated errors $\sigma^2_{spatial}$

An SSN model was fitted to the simulated data and the spatial covariance model extracted. This was then used to make random draws from a multivariate normal distribution with mean $\mathbf{0}$. Random draws were made

for 25 timepoints so each 'draw' consists of 30 values, one for each monitoring site. The draws at each time point should now be spatially correlated.

6 Temporally correlated errors $\sigma^2_{temporal}$

`arima.sim()` in `stats` package in R was used to produce errors with AR(1) temporal autocorrelation, where $\rho = 0.75$.

7 Independent errors σ^2_{iid}

Independent normally distributed errors were created using draws from a random normal distribution. Monitoring sites 2, 4, 7, 8, 11, 14, 18, 23, 25, 26, 27, 29 had an error term randomly drawn from a N(0,1) distribution. The remaining sites had an error term randoly drawn from a N(0,4) distribution.

8 Total error Σ_{total}

Total error, ε_{ij} where $i = 1, ..., 25$ time points and $j = 1, ..., 30$ monitoring sites was created by summing the spatial, temporal, and independent errors as

$$\Sigma_{total} = \sigma^2_{spatial} + \sigma^2_{temporal} + \sigma^2_{iid}$$

9 Adding a trend

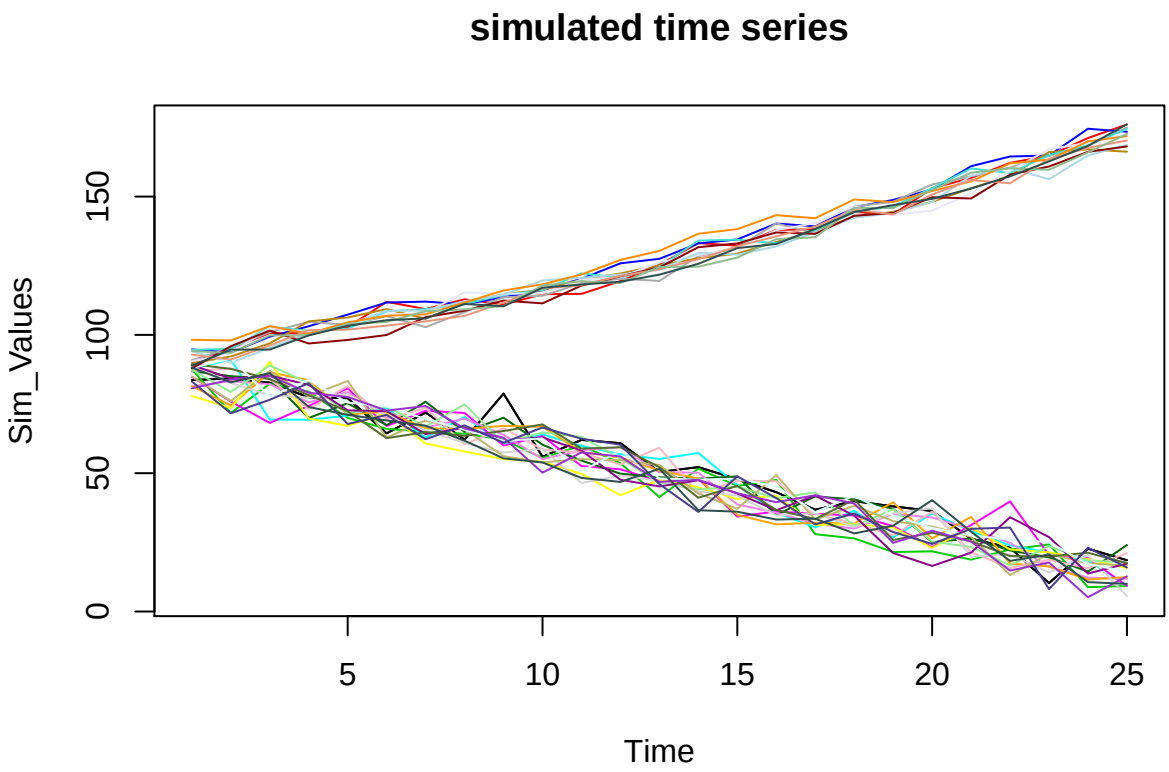
Two trends were created for the time series. Monitoring sites 2, 4, 7, 8, 11, 14, 18, 23, 25, 26, 27, 29 were allocated the following trend:

$$y_{ij} = \alpha_j + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_{ij}$$

where α_j represents the value simulated at each of $j = 1, ..., 30$ monitoring sites, $\beta_1 = 2$, $\beta_2 = 0.05$. For the rest of the monitoring sites, the trend was added as:

$$y_{ij} = \alpha_j + \beta_3 x_i + \varepsilon_{ij}$$

where $\beta_3 = -3$. The plot below shos the final time series with spatially and temporally correlated error, and two different temporal trends. A shift of 80 was added to all values.



These completed time series can be found in the stpca package in `demoY`.

10 Missing values

Missing values were created by randomly removing 15% (112) values from the demo data matrix, and the incomplete data can be found in `demoYmiss`.