The Journal of Open Source Software

SSN2: The next generation of spatial stream network modeling in R

Michael Dumelle[1], **Erin Peterson**[2], **Jay M. Ver Hoef**[3], **Alan Pearse**[4], **and Dan Isaak**[5]

1 United States Environmental Protection Agency **2** Queensland University of Technology **3** United States National Oceanic and Atmospheric Administration **4** University of Wollongong **5** United States Forest Service

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The `SSN2` **R** package provides tools for spatial statistical modeling and prediction on stream (river) networks. `SSN2` supersedes the `SSN` **R** package (J. Ver Hoef, Peterson, Clifford, & Shah, 2014), which was archived alongside broader changes in the **R**-spatial ecosystem (Nowosad, 2023) that included the retirement of `rgdal` (Bivand, Keitt, & Rowlingson, 2021), `rgeos` (Bivand & Rundel, 2020), and `maptools` (Bivand & Lewin-Koh, 2021). `SSN2` leverages modern **R**-spatial tools like `sf` (Pebesma, 2018) and provides many useful modeling features that were not feasible to implement in `SSN`.

Statement of Need

Streams (and rivers) are vital aquatic resources that sustain wildlife, provide drinking water, irrigate crops, and reduce pollution. Data are often collected at various locations on a stream network and used to characterize some aspect of the stream. For example, a researcher may be interested in how the amount of a hazardous chemical changes throughout the stream network and use this information to inform mitigation efforts. Spatial statistical models (i.e., geostatistical models) incorporate spatial covariance (i.e., dependence) among nearby observations to improve model fit and prediction accuracy (Cressie, 1993). Typically spatial covariance is modeled as a function of the Euclidean distance between observations. However, spatial covariance based solely on Euclidean distance fail to adequately describe unique and complex spatial dependencies on a stream network. Consider two pairs of sites that have the same stream distance between them but one pair shares flow among the sites and the other pair does not – these two pairs should have different amounts of covariance. The `SSN2` **R** package provides a formal structure for building stream network models that incorporates upstream processes (e.g., salmon swimming) and downstream processes (e.g., sediment transport).

The linear spatial stream network model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau}_{tu} + \boldsymbol{\tau}_{td} + \boldsymbol{\tau}_{eu} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is a matrix of explanatory variables (usually including a column of 1's for an intercept), $\boldsymbol{\beta}$ is a vector of fixed effects that describe the average impact of $\mathbf{X}$ on $\mathbf{y}$, $\boldsymbol{\tau}_{tu}$ is a vector of spatially dependent (correlated) tail-up random errors, $\boldsymbol{\tau}_{td}$ is a vector of spatially dependent (correlated) tail-down random errors, $\boldsymbol{\tau}_{eu}$ is a vector of spatially dependent (correlated) Euclidean random errors, and $\boldsymbol{\epsilon}$ is a vector of spatially independent (uncorrelated) random errors. The spatial dependence of each $\boldsymbol{\tau}$ term is explicitly specified using a spatial covariance function that incorporates the variance of the respective $\boldsymbol{\tau}$ term, often called a partial sill, and a range parameter that controls the behavior of the respective spatial covariance. The variance of $\boldsymbol{\epsilon}$ is often called the nugget (or nugget effect). The tail-up random errors incorporate dependence only among sites that share flow, while the tail-up random errors incorporate dependence among sites that share flow and do not share flow. The tail-up random errors also incorporate an "additive function" that reflects the upstream branching structure of the network (J. M. Ver Hoef & Peterson, 2010). J. M. Ver Hoef & Peterson (2010) and E. E. Peterson & Ver Hoef (2010) describe these models in more detail.

There are two main classes of functions in `SSN2`: 1) functions that operate on spatial stream network objects to either manipulate the SSN object, fit models, or simulate data (these have an `ssn_` prefix) and 2) functions that operate on a fitted model object, which are used to summarize the model, make predictions, and more.

Package Overview

Stream network data must be pre-processed using the STARS toolset for ArcGIS Desktop versions 9.3x - 10.8x (E. Peterson & Ver Hoef, 2014). STARS is used to create a `.ssn` folder, which contains all the topological information needed to fit stream network using `SSN2`. The `.ssn` folder is read into **R** using `ssn_import()`, which yields an SSN object that has a special list structure with four elements: `edges`, an `sf` object that contains network edges with `LINESTRING` geometry; `obs` an `sf` object that contains the observed data with `POINT` geometry; `preds`, a list of `sf` objects with `POINT` geometry that each contain a set of sites requiring prediction; and `path` a character string that stores the computer path to the `.ssn` object.

The `SSN2` packages comes with an example `.ssn` folder that represents temperature measurements from the Middle Fork stream network taken during 2004. To use this example data, we must copy the `.ssn` folder that comes shipped with `SSN2` to a temporary directory and store the temporary directory's file path:

```
library(SSN2)
copy_lsn_to_temp()
path <- paste0(tempdir(), "/MiddleFork04.ssn")
```

Copying to the temporary directory is only used in this illustrative example, as users will read in the `.ssn` object from an appropriate file path stored on their computer. Next we import the observed data and prediction sites (`pred1km`):

```
mf04p <- ssn_import(path, predpts = "pred1km")
```

We visualize the stream network, observed sties, and prediction sites using `ggplot2` (Wickham, 2016) by running

```
library(ggplot2)
ggplot() +
  geom_sf(data = mf04p$edges) +
  geom_sf(data = mf04p$preds$pred1km) +
  geom_sf(data = mf04p$obs, color = "brown", size = 1.5) +
  theme_bw()
```

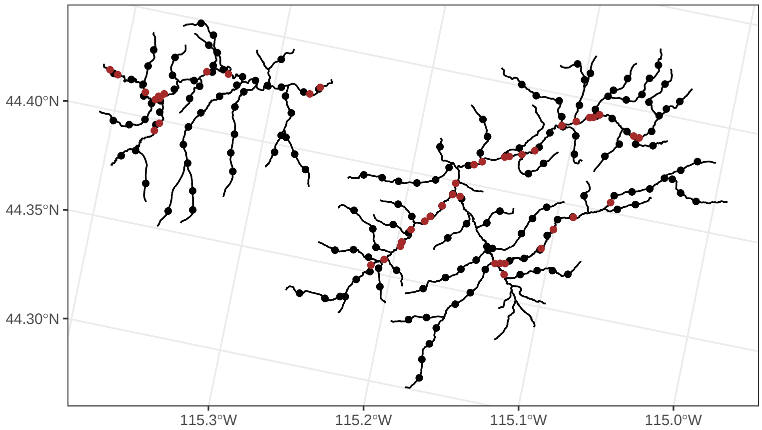

Figure 1: Middle Fork 2004 stream newtork. Observed sites are represnted by brown, closed circles. Prediction sites are represented by black, closed circles.

Suppose we wanted to summer model stream temperature (`Summer_mn`) as a function of elevation (`ELEV_DEM`) and precipitation (`AREAWTMAP`). We fit and summarize this model using `SSN2` by running

```
ssn_mod <- ssn_lm(
  formula = Summer_mn ~ ELEV_DEM + AREAWTMAP,
  ssn.object = mf04p,
  tailup_type = "exponential",
  taildown_type = "spherical",
  euclid_type = "gaussian",
  additive = "afvArea"
)
```

The `additive` argument represents a variable in `mf04p` that captures the "additive function value", which captures elements of the branching network used by the tail-up covariance (J. M. Ver Hoef & Peterson, 2010). A summary of the fitted model looks similar to a summary returned by `lm()` with information about the spatial covariance structure added:

```
summary(ssn_mod)
```

```
##
## Call:
## ssn_lm(formula = Summer_mn ~ ELEV_DEM + AREAWTMAP, ssn.object = mf04p,
##     tailup_type = "exponential", taildown_type = "spherical",
##     euclid_type = "gaussian", additive = "afvArea")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6393 -2.0646 -0.5952  0.2143  0.7497
##
## Coefficients (fixed):
##              Estimate Std. Error z value Pr(>|z|)
```

The Journal of Open Source Software

```
## (Intercept) 76.195041   7.871574   9.680  < 2e-16 ***
## ELEV_DEM    -0.026905   0.003646  -7.379  1.6e-13 ***
## AREAWTMAP   -0.009099   0.004461  -2.040   0.0414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Pseudo R-squared: 0.6124
##
## Coefficients (covariance):
##               Effect     Parameter   Estimate
##   tailup exponential  de (parsill)  3.800e+00
##   tailup exponential         range  4.194e+06
##   taildown spherical  de (parsill)  4.480e-01
##   taildown spherical         range  1.647e+05
##      euclid gaussian  de (parsill)  1.509e-02
##      euclid gaussian         range  4.496e+03
##               nugget        nugget  2.087e-02
```

SSN2 leverages the `tidy()`, `glance()`, and `augment()` functions (Robinson, Hayes, & Couch, 2021) to tidy, glance at, and augment the fitted model:

```r
tidy(ssn_mod, conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept) 76.2        7.87        9.68 0           60.8    91.6
## 2 AREAWTMAP   -0.00910    0.00446    -2.04 4.14e- 2   -0.0178 -0.000356
## 3 ELEV_DEM    -0.0269     0.00365    -7.38 1.60e-13   -0.0341 -0.0198
```

```r
glance(ssn_mod)
```

```
## # A tibble: 1 x 9
##       n     p  npar value   AIC  AICc logLik deviance pseudo.r.squared
##   <int> <dbl> <int> <dbl> <dbl> <dbl>  <dbl>    <dbl>            <dbl>
## 1    45     3     7  59.3  73.3  76.3  -29.6     41.9            0.612
```

```r
aug_mod <- augment(ssn_mod)
subset(aug_mod, select = c(Summer_mn, .fitted, .resid, .hat, .cooksd))
```

```
## Simple feature collection with 45 features and 5 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: -1530805 ymin: 2527111 xmax: -1503079 ymax: 2537823
## Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
## # A tibble: 45 x 6
##    Summer_mn .fitted .resid   .hat  .cooksd           geometry
##        <dbl>   <dbl>  <dbl>  <dbl>    <dbl>        <POINT [m]>
##  1      11.4    14.4  -3.07 0.0915 0.0962   (-1512690 2531883)
##  2      10.7    12.9  -2.20 0.114  0.00471  (-1512852 2531295)
##  3      10.4    12.7  -2.25 0.0372 0.00724  (-1513400 2530706)
##  4      10.1    12.3  -2.18 0.0251 0.00153  (-1514027 2530147)
```

```
## 5      10.1      12.3  -2.13 0.0374 0.000583 (-1514309 2529902)
## 6       9.81     12.0  -2.16 0.0602 0.0150   (-1515032 2529461)
## 7       9.76     11.6  -1.85 0.0736 0.00739  (-1515513 2528810)
## 8       9.77     11.6  -1.84 0.0648 0.00687  (-1515588 2528592)
## 9       9.53     11.4  -1.87 0.112  0.00152  (-1516440 2527899)
## 10     12.6      14.9  -2.28 0.0498 0.00964  (-1512464 2531195)
## # i 35 more rows
```

Specific helper functions (e.g., `coef()`, `AIC()`, `residuals()`) can be used to obtain the same quantities returned by `tidy()`, `glance()`, and `augment()`:

```
coef(ssn_mod)
```

```
## (Intercept)    ELEV_DEM    AREAWTMAP
## 76.19504087 -0.02690478 -0.00909941
```

```
AIC(ssn_mod)
```

```
## [1] 73.2623
```

```
head(residuals(ssn_mod))
```

```
##         1         2         3         4         5         6
## -3.066413 -2.204147 -2.252004 -2.175337 -2.131527 -2.162417
```

Prediction at the prediction sites is performed using `predict()` (or `augment()`):

```
aug_pred <- augment(ssn_mod, newdata = "pred1km", interval = "prediction")
subset(aug_pred, select = c(.fitted, .lower, .upper))
```

```
## Simple feature collection with 175 features and 3 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: -1530631 ymin: 2521707 xmax: -1500020 ymax: 2540253
## Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
## # A tibble: 175 x 4
##    .fitted .lower .upper          geometry
##      <dbl>  <dbl>  <dbl>       <POINT [m]>
##  1    14.6   14.3   15.0 (-1520657 2536657)
##  2    15.0   14.7   15.4 (-1519866 2536812)
##  3    14.8   14.3   15.3 (-1521823 2536911)
##  4    15.0   14.5   15.5 (-1523183 2537256)
##  5    15.2   14.7   15.6 (-1523860 2537452)
##  6    15.1   14.8   15.5 (-1525443 2537698)
##  7    15.1   14.7   15.5 (-1526397 2537254)
##  8    15.0   14.6   15.4 (-1527436 2536803)
##  9    14.9   14.6   15.3 (-1529043 2536449)
## 10    14.9   14.5   15.2 (-1529689 2537313)
## # i 165 more rows
```

Generalized linear models for binary, count, proportion, and skewed data are available via the `ssn_glm()` function. Simulating data on a stream network is performed via `ssn_simulate()`.

Discussion

After the retirement of `SSN`, `SSN2` is the primary tool available in the **R** ecosystem available to fit models, make predictions, and simulate data on spatial stream networks. Spatial stream network models are incredibly useful and have motivated research studying water quality (Mainali, Chang, & Chun, 2019), fish populations (Isaak, Ver Hoef, Peterson, Horan, & Nagel, 2017), temperature patterns (Detenbeck, Morrison, Abele, & Kopp, 2016), and salinization (McManus et al., 2020), among many other applications (e.g., Isaak et al. (2014); Neill et al. (2018); Pearse et al. (2020); Fuller et al. (2021); Santos-Fernandez et al. (2022)).

Acknowledgements

The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of USEPA, NOAA, or USFS. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government, USEPA, NOAA, or USFS. USEPA, NOAA, or USFS do not endorse any commercial products, services or enterprises.

References

Bivand, R., Keitt, T., & Rowlingson, B. (2021). *Rgdal: Bindings for the 'geospatial' data abstraction library*. Retrieved from https://CRAN.R-project.org/package=rgdal

Bivand, R., & Lewin-Koh, N. (2021). *Maptools: Tools for handling spatial objects*. Retrieved from https://CRAN.R-project.org/package=maptools

Bivand, R., & Rundel, C. (2020). *Rgeos: Interface to geometry engine - open source ('GEOS')*. Retrieved from https://CRAN.R-project.org/package=rgeos

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Detenbeck, N. E., Morrison, A. C., Abele, R. W., & Kopp, D. A. (2016). Spatial statistical network models for stream and river temperature in new england, USA. *Water Resources Research*, *52*(8), 6018–6040.

Fuller, M. R., Ebersole, J. L., Detenbeck, N. E., Labiosa, R., Leinenbach, P., & Torgersen, C. E. (2021). Integrating thermal infrared stream temperature imagery and spatial stream network models to understand natural spatial thermal variability in streams. *Journal of thermal biology*, *100*, 103028.

Isaak, D. J., Peterson, E. E., Ver Hoef, J. M., Wenger, S. J., Falke, J. A., Torgersen, C. E., Sowder, C., et al. (2014). Applications of spatial statistical network models to stream data. *Wiley Interdisciplinary Reviews: Water*, *1*(3), 277–294.

Isaak, D. J., Ver Hoef, J. M., Peterson, E. E., Horan, D. L., & Nagel, D. E. (2017). Scalable population estimates using spatial-stream-network (SSN) models, fish density surveys, and national geospatial database frameworks for streams. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(2), 147–156.

Mainali, J., Chang, H., & Chun, Y. (2019). A review of spatial statistical approaches to modeling water quality. *Progress in Physical Geography: Earth and Environment*, *43*(6), 801–826.

McManus, M. G., D'Amico, E., Smith, E. M., Polinsky, R., Ackerman, J., & Tyler, K. (2020). Variation in stream network relationships and geospatial predictions of watershed conductivity. *Freshwater Science*, *39*(4), 704–721.

Neill, A. J., Tetzlaff, D., Strachan, N. J. C., Hough, R. L., Avery, L. M., Watson, H., & Soulsby, C. (2018). Using spatial-stream-network models and long-term data to understand and predict dynamics of faecal contamination in a mixed land-use catchment. *Science of the Total Environment*, *612*, 840–852.

Nowosad, J. (2023, June 4). Upcoming Changes to Popular R Packages for Spatial Data: What You Need to Do. Retrieved from https://geocompx.org//post/2023/rgdal-retirement

Pearse, A. R., McGree, J. M., Som, N. A., Leigh, C., Maxwell, P., Ver Hoef, J. M., & Peterson, E. E. (2020). SSNdesign—an r package for pseudo-bayesian optimal and adaptive sampling designs on stream networks. *PloS one*, *15*(9), e0238422.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. doi:10.32614/RJ-2018-009

Peterson, E. E., & Ver Hoef, J. M. (2010). A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology*, *91*(3), 644–651.

Peterson, E., & Ver Hoef, J. (2014). STARS: An ArcGIS toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. *Journal of Statistical Software*, *56*, 1–17.

Robinson, D., Hayes, A., & Couch, S. (2021). *Broom: Convert statistical objects into tidy tibbles*. Retrieved from https://CRAN.R-project.org/package=broom

Santos-Fernandez, E., Ver Hoef, J. M., Peterson, E. E., McGree, J., Isaak, D. J., & Mengersen, K. (2022). Bayesian spatio-temporal models for stream networks. *Computational Statistics & Data Analysis*, *170*, 107446.

Ver Hoef, J. M., & Peterson, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, *105*(489), 6–18.

Ver Hoef, J., Peterson, E., Clifford, D., & Shah, R. (2014). SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, *56*, 1–45.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org