

K1D: Multivariate Ripley's K-function for one-dimensional data

Daniel G. Gavin
University of Oregon
Department of Geography
Version 1.1 (May 2007)

Contents

1. Background

- 1a. Bivariate and multivariate K-Function for one dimension
- 1b. Use of integer data
- 1c. Confidence envelope for the multivariate K-function
- 1d. Smoothing event frequencies

2. Example of the bivariate K-function for one dimensional data

3. Running K1D

Revision History:

Revision Beta 2 (August 2005). Fixed file input error. Now does not require the first column to have the greatest number of events.

Revision Beta 3 (November 2005). Added several options: different model forms (forward and backward selection), use of integer data, and different randomization schemes for constructing confidence envelopes.

Version 1.0 (February 2007). Changed format of input file to include specifications of start-year and end-year. Otherwise, no other changes.

Version 1.1 (May 2007). Fixed error in calculation of the forward-selection and backward-selection functions.

1. Background

K1D computes the multivariate Ripley K -function simplified for one dimension (e.g., time or line transects). K1D computes the dependence between two or more types of events that are ordered in one dimension. The motivation for creating this program was to test whether dates of forest fires at two or more sites co-occur more than would be expected by chance. A general review of the K -function and the bivariate K -function will not be covered here, but an excellent recent review is in Wiegand and Moloney (2004).

K1D has three main features:

- a) Calculation of the bivariate K -function and its transform (L) using an edge correction. K function may be calculated one of three ways.
- b) Calculation of confidence envelopes one of three ways: 1) randomization of events (optionally weighted by an intensity function), 2) a ‘circular’ (‘toroidal’ in two dimensions) shift of the records relative to each other, or 3) a shuffle of events (i.e., randomization without replacement).
- c) Calculation of smoothed frequencies of events using a set window width.

1a. Bivariate and multivariate K -Function for one dimension

The bivariate K -function gives the number of events in record B occurring within $\pm t$ yr (the ‘temporal window’) of each event in record A , and scaled by $T/(n_A n_B)$ where T is the length of the record (transect) and n_A and n_B are the number of events in A and B respectively. The bivariate K -function for a single dimension is expressed as:

$$\tilde{K}_{AB}(t) = \frac{T}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} w(A_i, B_j) \mathbf{I}(|A_i - B_j| < t)$$

where A_i and B_j are locations (times) of events, \mathbf{I} is the identity function, and $w(A_i, B_j)$ is an edge correction, set to 2 if $|A_i - B_j|$ is greater than the distance of A_i to the nearest ‘edge’ of the record, otherwise it is set to 1 (Doss 1989). Values of K greater than $2t$ suggest attraction, or synchrony, between A and B , within a window of t . Values of K close to $2t$ suggest no relationship or independence between A and B , and values less than $2t$ suggest repulsion, or asynchrony. The edge correction causes $\tilde{K}_{AB}(t)$ to differ slightly from $\tilde{K}_{BA}(t)$ (i.e., whether distances are measured from A to B or vice versa). $\tilde{K}_{AB}(t)$ and $\tilde{K}_{BA}(t)$ are averaged by weighting each to produce the unbiased estimate $\hat{K}_{AB}(t)$ (Lotwick and Silverman 1982):

$$\hat{K}_{AB}(t) = \frac{n_B \cdot \tilde{K}_{AB}(t) + n_A \cdot \tilde{K}_{BA}(t)}{n_A + n_B}$$

The multivariate expression of the K -function (comparing ≥ 2 types of events) is implemented by comparing events one record to the aggregated events in all other records:

$$K_{CG}(t) = \frac{T}{\sum_{R=1(R \neq C)}^G n_C n_R} \sum_{i=1}^{n_C} \sum_{R(R \neq C)}^G \sum_{j=1}^{n_R} w(C_i, R_j) \mathbb{I}(|C_i - R_j| \leq t)$$

where G is the number of records and C is a single record compared to all other records. The above function is calculated for all G records, and then the G functions are averaged as for the bivariate case:

$$K_{GG}(t) = \frac{\sum_{C=1}^G \left(\sum_{R(R \neq C)}^G n_R \cdot K_{CG}(t) \right)}{\sum_{R=1}^G n_R (G-1)}$$

For graphing purposes, the K -function is easier to interpret if its mean and variance are stabilized over t as expressed by the L -function: $\hat{L}_{AB}(t) = \hat{K}_{AB}(t)/2 - t$.

The above methods are suitable for examining whether events occur closer to events in a second record irrespective of direction. In some cases, one may wish to focus on whether events in one record follow events in another record. This is termed the “forward selection” K -function:

$$\tilde{K}_{AB}(t) = \frac{T}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \mathbb{I}((B_j - A_i) \leq t, B_j \geq A_i)$$

where the K -function is the same as the previous bi-directional model, but there is an added criteria that ‘B’ events co-occur with or follow ‘A’ events (see after the comma in the identify function).

Similarly, one can define the opposite case, i.e., whether events in one record precede events in another record. This is the ‘backward selection’ K -function for ‘B’ events preceding ‘A’ events:

$$\tilde{K}_{AB}(t) = \frac{T}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \mathbb{I}((A_i - B_j) \leq t, A_i \geq B_j)$$

Because events are tallied in only one direction, the L -function for both forward and backward selection models is $\hat{L}_{AB}(t) = \hat{K}_{AB}(t) - t$.

Notes on implementation in KID: In this software, edge corrections are not applied with the ‘forward selection’ or ‘backward selection’ methods. The program reads the first column of data as the ‘A’ record, and the second column as the ‘B’ record. Thus, for ‘forward selection’, the K -function searches for ‘B’ events following ‘A’ events. For the ‘backward selection’ method, the K -function

searches for ‘B’ events preceding ‘A’ events. Only two records are permitted with the forward and backward selection methods.

1b. Use of integer data

In many cases, one wishes to compare series using values at a resolution of whole numbers. For example, time series often have only one value per year, and events cannot be defined at resolutions less than a year. In this case, it is possible to specify integer data for calculation of the K-function and for the simulation envelopes. For integer data, K1D will compute a K-function using $t=0$ in addition to the increments specified. It is necessary that the increment of the K-function is an integer when using integer data.

1c. Confidence envelope for the multivariate K-function

A confidence envelope is constructed using simulations. There are three simulation methods:

1. ‘Circular’ shift: Each randomization consists of shifting each record a random number of years and wrapping events from the end to the start of the record. This test only addresses dependence between the two records and preserves the first-order properties (frequency) of each record in each randomization (Wiegand and Moloney 2004). Confidence envelopes are set by percentiles in the distribution of the simulated K-function. Where $\hat{L}_{AB}(t)$ is greater than, within, or less than the confidence envelope indicates statistically significant synchrony (attraction), independence, or statistically significant asynchrony (repulsion), respectively, in a window of t .
2. Random numbers: Events are chosen randomly (a Poisson process). Events may also be weighted if there is a non-stationary trend that must be preserved in the simulations. Such weightings may be determined using Poisson regression.
3. Shuffling of events. This is only available for integer data. Events are chosen randomly without replacement.

1d. Smoothing event frequencies

To help visualize changes in event frequencies using different levels of smoothing, K1D also computes smoothed event frequencies. This smoothing is done in two steps. First, along regular intervals $l \in y, 2y, 3y, 4y \dots T$, number of events occurring within a window of $\pm w$ is tallied. This number is scaled upward for cases where the window extends beyond the edge of the record:

$$F(l) = \sum_{i=1}^n \begin{cases} 1, & \min(X_i, T - X_i) \geq w \\ \frac{w}{\min(X_i, T - X_i) + 0.5w}, & \min(X_i, T - X_i) < w \end{cases}$$

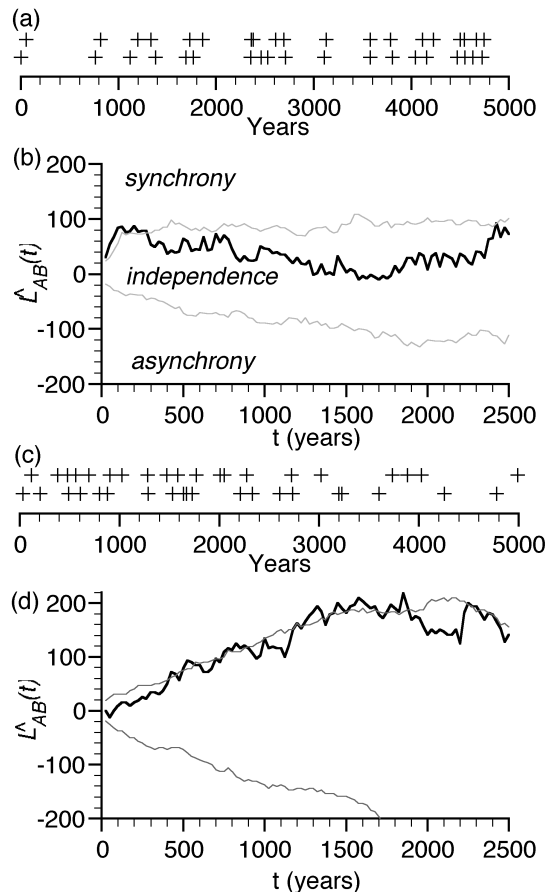
This results in a step-like series of event frequencies. Second, $F(l)$ is smoothed using the tricube function using the same window width following Huntley et al. (1989):

$$S(l) = \frac{\sum_{m=1: (|ym-l|<w)}^{T/y} V(l,ym) \cdot F(ym)}{\sum_{m=1: (|ym-l|<w)}^{T/y} V(l,ym)}$$

where $S(l)$ is the smoothed value at position l with a window width of $\pm w$, m the step number of size y such that ym is the position the position along the record, and $V(l,ym)$ is the tricube weighting function. Smaller y results in a less jagged curve and $y \ll w$. The tricube function for the weighting of the distance between l and ym is:

$$V(l,ym) = \left(1 - \left(\frac{|ym-l|}{w} \right)^3 \right)^3$$

2. Example of the bivariate K-function for one dimensional data



Demonstration of the bivariate K -function for testing synchrony over a range of temporal windows. (a) Two records where the first is a series of events placed on random years and the second has events placed within 50 yr of events in the first record. (b) The L -function (transform of the K -function) for the events in (a) with 95% confidence envelope (thin lines) based on 1000 randomizations. The function exceeds the upper confidence envelope from 25 to 150 yr, indicating strong correlation of event times within windows of that scale, but lack of long-term patterns in both records results in no synchrony in larger windows. (c) Two records where the number of events decreases by one each millennium. (d) The L -function for the events in (c). The function exceeds the confidence envelope at several window sizes between 500 and 1700 yr, indicating the millennial-scale pattern in common between the two records.

3. Running K1D

This program is compiled for Mac OS X and Windows. The Windows version has received little testing (version Beta 3).

Data input: Tab-delimited text files

5	0	2000		
CY	MIR	RS	Yahoo	Barr
349	136	403	24	1360
445	282	815	907	1530
569	410	915	1267	1600
1004	1073	1034	1928	1640
1099	1288	1240		
1313	1426	1427		
1485	1605	1595		
1801	1746	1695		
		1777		
		1857		

The first row contains the number of records, the start year, and the end year. Here, the records span from 1-2000. The length of the record (T) is the difference of the start year and end year. **All events must be bounded by the start-year and end-year dates. If using integer data, the data are inclusive of the start year and end year, and thus the length of the record is endyear-startyear+1.**

The second row contains the names of each record.

All other rows contain the locations of events. Integer or non-integer values.

- 1) Load files using File...Open or Command-O. If the files are read properly, a text box on the program will print the number of events in each record.
- 2) Choose the model form with the pop-up menu.
- 3) Choose the number of intervals in the K-function. The K-function will be computed to $T/2$, where T is the length of the record. For example, 10 intervals on a 2000-year record will yield K computed from 100 to 1000 at steps of 100 units. Run the K -function.
- 4) Choose the method of randomizing, and the records to randomize. For the 'random' method, you may optionally choose an intensity-weighting file to bias event positions in simulated records. The intensity weighting file is a tab-delimited text file with two columns: column 1=location in record; column 2=probability occurring at that point. Column 1 should have rows for 1 to the length of the record. Column 2 should sum to 1 over all rows. There is a header row; see example file. The scale must be such that the intensity function can be described for the entire length of the record using integers (i.e., a record of length 500 is better than a record of length 3). This limitation may be addressed in future versions.
- 5) Choose the confidence envelope interval and the number of simulations to run. Larger numbers of simulations produces more reliable confidence envelopes; usually 1000 is sufficient. Run the simulations. A progress bar will show how much time remains for these calculations.
- 6) Output, with a header row, is copied using Command-C, and can be pasted into spreadsheets. Do not click on the spreadsheet before copying, as only the selected cells will be copied.
- 7) Run the smoothing by selecting a window width and the step size for the output. The window width is the total width of the window: a window of 100 results in a binning of events occurring ± 50 from a regularly stepped position. The step size should be smaller than the window width, and results in a more appealing curve if 100 or more steps are used.
- 8) New data files may be opened without quitting and restarting.

References

- Doss, H. 1989. On estimating the dependence between two point processes. *The Annals of Statistics* **17**:749-763.
- Huntley, B., Bartlein, P.J., & Prentice, I.C. (1989) Climatic control of the distribution and abundance of beech (*Fagus L*) in Europe and North America. *Journal of Biogeography*, **16**, 551-560.
- Lotwick, H. W., and B. W. Silverman. 1982. Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society* **B44**:406-413.
- Wiegand, T., and K. A. Moloney. 2004. Rings, circles, and null-models for point pattern analysis in ecology. *Oikos* **104**:209-229.