Goodness of fit testing in RSiena

Tom A.B. Snijders

University of Groningen
University of Oxford

February, 2023

Two functions are available in RSiena
for checking model assumptions:

1. <sienaTimeTest>
   for testing time homogeneity
   (meaningful only if there are 3 or more waves);

Two functions are available in RSiena
for checking model assumptions:

1. <sienaTimeTest>
   for testing time homogeneity
   (meaningful only if there are 3 or more waves);
2. <sienaGOF>
   for checking that the RSiena model reproduces sufficiently
   the characteristics of the observed networks.

Both were initially developed by Josh Lospinoso (Oxford).

.

sienaTimeTest

For M waves there are $M-1$ periods.

The assumption
that parameters are constant in the $M-1$ periods
is tested by <sienaTimeTest>.

The <summary> method also produces effect-wise
and period-wise tests.

See `RscriptSienaTimeTest.r`

sienaTimeTest

For M waves there are $M-1$ periods.

The assumption
that parameters are constant in the $M-1$ periods
is tested by <sienaTimeTest>.

The <summary> method also produces effect-wise
and period-wise tests.

See `RscriptSienaTimeTest.r`

Can be used also for multi-group projects;
then, homogeneity of the groups is tested.

.

The associated function <includeTimeDummy> can be used to interact the effects specified by time dummies, representing time heterogeneity.

An alternative for this purpose is to define time variables (dummies or trend or other time-dependent variables) and add those to the data set, and then specify interactions between the other effects and these time variables.

This is a bit more work but also more flexible and clearer.

E.g., you can use time dummies that are 0 first and 1 later on, and stay equal to 1.

Or you could use a linear, quadratic, logarithmic, ..., trend.

.

sienaGOF

A good model for network dynamics should represent the important features of the network in a good way.

For the features that are expressed by the target statistics, directly corresponding to the effects included in the model, this is guaranteed by adequate convergence of the MoM.

However, also other features should be represented well.
For example, the distribution of indegrees and outdegrees, and the triad census.
Perhaps also the distribution of geodesic distances.

This was brought forward as a goodness-of-fit criterion for ERGMS by Hunter, Goodreau, and Handcock (*JASA*, 2008).

.

For SAOMs, the goodness of fit of a model
can be tested by the function <sienaGOF>.

See Lospinoso & Snijders (*Methodological Innovations*, 2019).

This requires that <siena07> was run
with `returnDeps = TRUE`.

This option returns the simulated data sets in Phase 3
as part of the <sienaFit> object produced by <siena07>.

.

(from the help page ...:)

This is done by simulations of auxiliary statistics,
different from the statistics used for parameter estimation.
The fit is good if the average values
of the auxiliary statistics over many simulation runs
are close to the values observed in the data.

A Monte Carlo test based on the Mahalanobis distance
is used to calculate *p*-values.

This is a case where you wish the *p*-values to be *large* enough!

A <plot> method can be used to diagnose poor fit.

.

The auxiliary statistics must be given explicitly
in the call of <sienaGOF>.

Some basic auxiliary statistics are available directly:

<OutdegreeDistribution>
<IndegreeDistribution>
<BehaviorDistribution>
<TriadCensus>
<mixedTriadCensus>
<dyadicCov> ;

and the user can also create custom functions.

The help page <sienaGOF-auxiliary> contains
some additional functions using packages `igraph` and `sna`.

.

Sketch of the use of sienaGOF

See `?sienaGOF` and the script `sienaGOF_vdB.R`

The basic operation is as follows:

```
results1 <- siena07(myalg, data=mydata, effects=myeff,
                    returnDeps=TRUE)
gof1.od <- sienaGOF(results1, verbose=TRUE,
                    varName="friendship", OutdegreeDistribution,
                    cumulative=TRUE, levls=0:10)
gof1.od
plot(gof1.od)
```

You can adapt the parameters `levls` and `cumulative` in <sienaGOF>.

Auxiliary functions

Some auxiliary functions are available within RSiena,
('out of the box'),
some are listed on the help page for <"sienaGOF-auxiliary">,
such as <GeodesicDistribution>,
and others can be made by yourself (...)
or in future by others (!!!).

If you wish to use <GeodesicDistribution>,
you have to take this function
from the <sienaGOF-auxiliary> help page and give it to R.

What is available now is not meant to be complete!

.

How good a fit is required?

Since some time we have been moving to a new standard for publications using Siena, where the fit for the degree and behavior distributions should be adequate.

Of course it is also advisable to consider goodness of fit for the triad census and the geodesic distribution.

It may not always be possible to achieve a fit with $p > 0.05$ for the Mahalanobis combination of all statistics under consideration.

But it should be attempted, and in my experience it usually is possible, to have the data within the confidence band of *plot.sienaGOF* for the degree and behavior distributions.

How to specify the model?

This depends of course on the purpose of the research, theoretical considerations, empirical knowledge...

But the following may be a guideline
for specifying the network model:

1. Outdegree effect: always.
2. Reciprocity effect: almost always.
3. A triadic effect representing network closure.
   gwesp, transitive triplets, and/or transitive triads.
4. transitive reciprocated triplets and/or three-cycles
   (see Block, *Social Networks*, 2015).

.

How to specify the model? *(continued)*

5. Degree-related effects:
   indegree-popularity ('Matthew effect'), outdegree-activity,
   outdegree-popularity and/or indegree-activity
   perhaps reciprocated degree-activity;
   (raw or sqrt versions depending on goodness of fit).
6. Think about what are important covariates!
7. For networks of size larger than (say) 50:
   there may be institutional or contextual constraints /
   opportunities determining contact opportunities;
   distance (logged), same classroom, etc.
8. For binary actor covariates: ego, alter, same effect.

.

How to specify the model? *(continued further)*

9. For numerical actor covariates:
   think of the five-parameter model
   (Snijders & Lomi, *Network Science* 2019)
   ego, ego squared, alter, alter squared,
   difference squared or ego × alter effect;
   for important covariates: check all five,
   reduce depending on empirical results.

10. If there is a strong center-periphery structure,
    and/or a strong dispersion in the outdegrees,
    then a dependence of the rate function on the
    log-outdegree (*outRateLog*) may be advisable.

.

Another guideline is that the model should allow you to answer your research questions (of course), and it should also have a good fit to the data.

A large set of effects is available in RSiena, growing over the years because of researchers' requests.

.

The fit can be checked, to some extent,
by using <sienaTimeTest> and <sienaGOF>.

11. If the fit for the degree distribution is poor,
you may try additional degree-related effects (isolation!).
12. If the fit for triad census or geodesic distribution is poor,
you may try additional triadic effects.

However, in both cases, THINK!!!

Are you missing additional covariate or structural effects?
Think of interactions, covariate transformations!

Auxiliary function <dyadicCov> for <sienaGOF> can be used
to check fit for ego-alter combinations of monadic variables;
see help page for <sienaGOF-auxiliary> (from version 1.2-25).

.

Note that difficulties in obtaining
convergence of the estimation procedure
may be a sign
of model misspecification or overspecification.

(The converse is not true!!!)

.

Further remarks

See the help pages for further information,
and Sections 5.11, 5.13, and 8.6 of the manual.

Also see the scripts on the Siena scripts webpage
and help page for <sienaGOF-auxiliary>.

First test time homogeneity, then goodness of fit.

Goodness of fit testing can be time consuming;
you may explore it with a Phase 3 of reduced length.

Testing of time homogeneity and goodness of fit
is starting to become more and more important.

The <summary> method of <sienaGOF> will give rough estimates
of the improvement in fit when parameters specified with
`fix=TRUE, test=TRUE` would be set free.