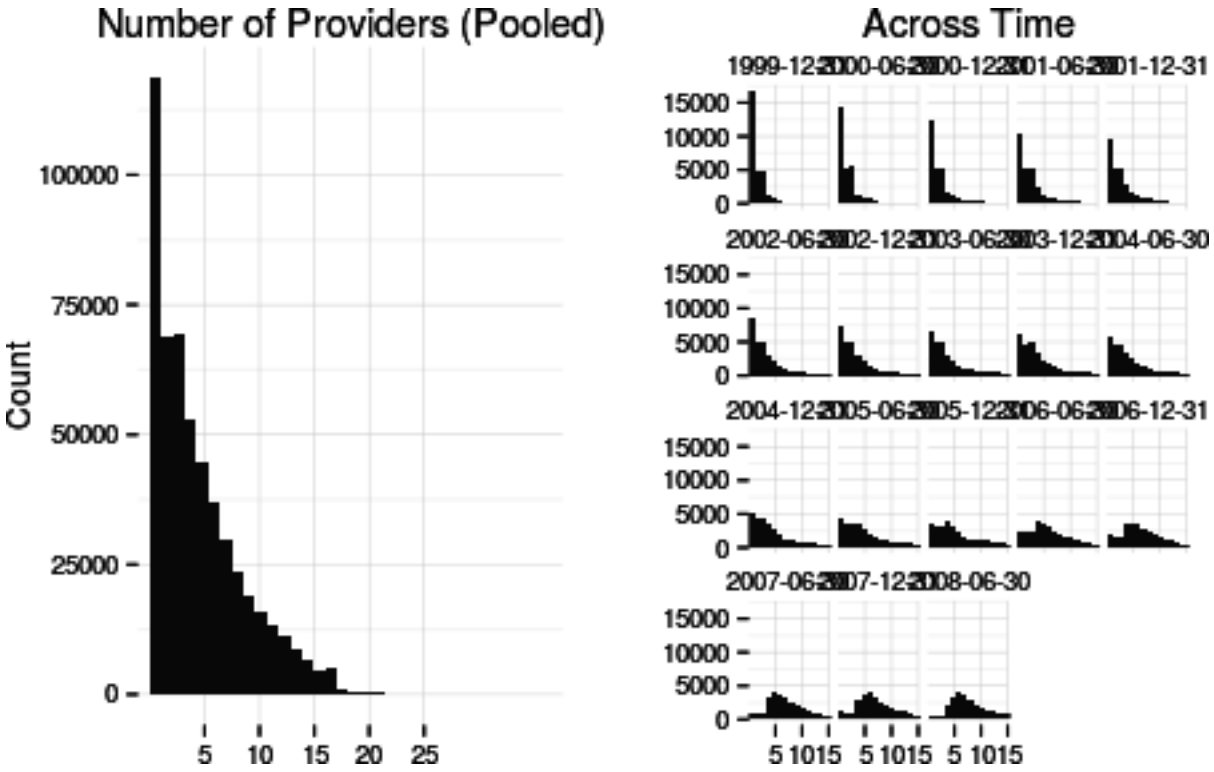
4-USDA Evaluation Poisson

Robert Dinterman

2015-10-22

Poisson Panel Regressions for Broadband Availability

The following regressions make use of the `glm` function for generalized linear models in `R` in order to identify the relationship between broadband availability and the USDA Broadband Loan Program. Loosely, we wish to model the number of broadband providers by zip code across the years 1999 to 2008 and determine whether or not the USDA Broadband Loan Program had an impact on broadband availability as this was one of the intended benefits of the subsidized loans. As a quick reference, here is what the dependent variable looks like across time:



Suppressed values were drawn from a uniform distribution between 1 and 3 for visual purposes.

Data on broadband providers is measured twice a year (June 30 and December 31) and takes on a count value of 0, 1-3*, 4, 5, 6, ... 31. The value 1-3* is a suppressed value of broadband providers for confidentiality purposes and has been coded as 2 to be consistent with the literature. In the sample, the mean across all years for providers is 4.7 and a sample variance of 15. For a Poisson distribution, the mean and variance are restricted to be equal which indicates here that the unconditional distribution is not likely to be Poisson. This may be troublesome if the conditional mean and variance of the model differ by as much as the sample mean and variance. Therefore, Quasi-Poisson and Negative Binomial models are also considered (and likely to be the true process).

Other variables used include:

- `iloans` - an indicator variable for whether or not a zip code received a loan and the central focus. No loans were given before 2002, so this variable does vary by time. This can be borken down further to

`ipilot` to indicate the smaller Pilot program which lasted from 2001 to 2003 and disbursed 28 loans totalling $180 million as well as the Post-Pilot program, `ibip1234`, which was established by the 2002 Farm Bill. The Post-Pilot program began disbursing loans in 2003 and we have data until 2006 which totals $1.22 billion across 70 loans.

- `log(est)` - this is from zip code business pattern data and is the number of establishments in a particular zip code. I take the log of this variable because the distribution is right-skewed. It is possible to substitute this variable with number of employees, annual payroll, or first quarter payroll from the ZBP but I choose not to because those variables are **suppressed for approximately 9% of the zip codes**. Establishments is highly correlated with the other variables anyway, so I would rather use a less precise proxy than potentially bias the sample.
- `log(Pop_IRS)` - IRS has data on number of tax returns filed by county from 1989 until 2013. I use the number of exemptions per county as a way to proxy for the population of a county. This variable is also right-skewed and therefore the log of population is taken instead of population. The alternative for population would be to use US Census data which produce yearly estimates at the county level. These estimates are based off of the 2000 Census and use the demographic age distribution of a county in order to project forward the birth rate and death rate to determine what the population in a county should be. Since this is simply a function of initial conditions in 2000, I choose to use IRS data because there is more variation in the data and it reflects changes in economic conditions across counties that would drive migration (population change).
- `logINC` - this is tabulated from the same IRS data above using Adjusted Gross Income (AGI) at the county level for each year. This is divided by the number of households for a county and is therefore a proxy for mean income, as reported to the IRS, per year. Again, this is a right-skewed variable which is the justification for taking the logarithm of the variable.
- `tri` - stands for Terrain Ruggedness Index which uses elevation data for a given polygon to calculate the feature changes in a given area relative to the entire domain. This is at the ZCTA level across the United States and is thought of as a proxy for increased costs of broadband deployment due to rough terrain. This does not vary across years and so zip code fixed effects will take away this variable.
- `ruc` - the rural-urban continuum code, but for this study I simply use 3 classifications of a county: Metro, Rural but adjacent to a metro county, and Rural but non-adjacent to a metro county. Counties do change across time, but only in years that end in 3 (1993, 2003, ...). I choose to use the values for 2003 as this would be a little bit before the halfway point in the analysis.

Poisson Regression Models

I start with by making use of the count nature of the broadband providers variable by assuming it follows a Poisson distribution:

$$Prov_{z,t}|X_{z,t}, \beta \sim Pois\left(\lambda_{z,t}\right)$$
$$log\left(\lambda_{z,t}\right) = \beta_0 + \beta_1 Loan_{z,t} + \beta_2 X_{z,t} + \tau_t + \varepsilon_{z,t}$$

The variable $Prov_{z,t}$ is the number of providers in zip code z at time t. The $X_{z,t}$ are variables at the zip code or county level that determine the level of broadband providers. These are log of establishments, log of population, log of income, terrain ruggedness index, and rural: adjacent and non-adjacent. There is also a time fixed effect for each year included in these regressions. The biannual values for the provider numbers gives a panel dataset where T = 18 and n = 29588.

Biannual Regression

As a naive start to use of the Poisson distribution, I will start with two models: the first without time fixed effects and the second including these. An ANOVA test to determine whether the time fixed effects are jointly significant is performed at the bottom which provides evidence of time fixed effects as a significant predictor of broadband diffusion:

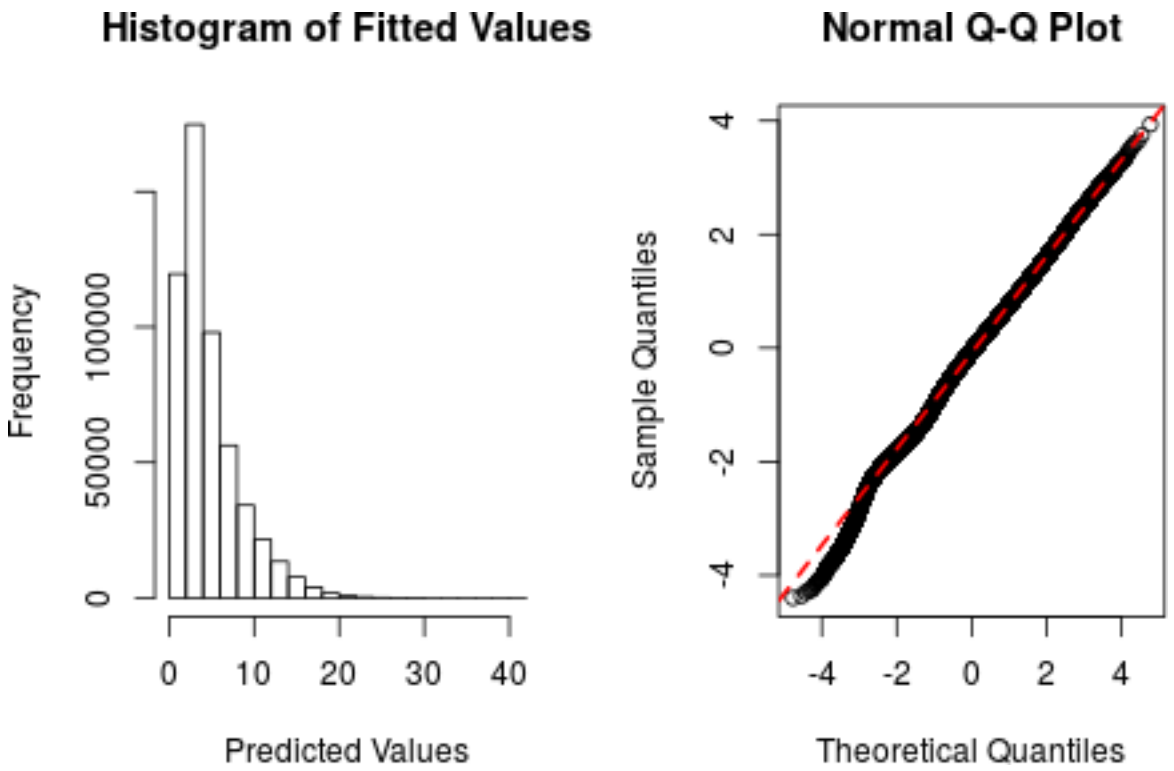
```
##
## Call:
## glm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta),
##     family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6157  -0.9495  -0.1428   0.7024   4.9423
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.105e-01  3.050e-02  -6.901 5.17e-12 ***
## iloans               3.275e-01  3.548e-03  92.316  < 2e-16 ***
## log(est)             2.569e-01  4.876e-04 526.827  < 2e-16 ***
## log(Pop_IRS)         9.038e-02  5.950e-04 151.898  < 2e-16 ***
## logINC              -4.481e-02  2.968e-03 -15.098  < 2e-16 ***
## tri                 -8.285e-04  3.662e-05 -22.626  < 2e-16 ***
## rucadj              -3.109e-02  2.151e-03 -14.455  < 2e-16 ***
## rucnonadj           -4.242e-02  2.648e-03 -16.016  < 2e-16 ***
## poly(AREA_zcta, 2)1 -5.103e+00  5.810e-01  -8.784  < 2e-16 ***
## poly(AREA_zcta, 2)2  3.183e+00  5.247e-01   6.066 1.31e-09 ***
## I(Pop_IRS/AREA_cty)  4.581e-06  2.118e-07  21.625  < 2e-16 ***
## I(est/AREA_zcta)     2.103e-06  9.127e-07   2.304   0.0212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1668357  on 532583  degrees of freedom
## Residual deviance:  825770  on 532572  degrees of freedom
## AIC: 2414377
##
## Number of Fisher Scoring iterations: 5


##
## Call:
## glm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) +
##     factor(time), family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3988  -0.6521  -0.0568   0.4923   3.9361
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.952e+00  3.104e-02 -95.095  < 2e-16 ***
## iloans                 7.863e-03  3.578e-03   2.198   0.0280 *
## log(est)               2.503e-01  4.875e-04 513.430  < 2e-16 ***
## log(Pop_IRS)           7.518e-02  5.980e-04 125.710  < 2e-16 ***
## logINC                 1.286e-01  2.971e-03  43.273  < 2e-16 ***
## tri                   -1.249e-03  3.681e-05 -33.932  < 2e-16 ***
## rucadj                -1.897e-02  2.155e-03  -8.803  < 2e-16 ***
```

```
## rucnonadj              -3.547e-02  2.653e-03 -13.371  < 2e-16 ***
## poly(AREA_zcta, 2)1    -3.332e+00  5.786e-01  -5.760 8.43e-09 ***
## poly(AREA_zcta, 2)2     1.390e+00  5.502e-01   2.527   0.0115 *
## I(Pop_IRS/AREA_cty)     4.201e-06  2.111e-07  19.903  < 2e-16 ***
## I(est/AREA_zcta)        1.668e-06  9.136e-07   1.825   0.0680 .
## factor(time)2000-06-30  2.185e-01  6.354e-03  34.387  < 2e-16 ***
## factor(time)2000-12-31  4.762e-01  6.025e-03  79.045  < 2e-16 ***
## factor(time)2001-06-30  6.522e-01  5.841e-03 111.655  < 2e-16 ***
## factor(time)2001-12-31  7.172e-01  5.778e-03 124.133  < 2e-16 ***
## factor(time)2002-06-30  8.397e-01  5.667e-03 148.189  < 2e-16 ***
## factor(time)2002-12-31  9.202e-01  5.600e-03 164.312  < 2e-16 ***
## factor(time)2003-06-30  1.014e+00  5.527e-03 183.493  < 2e-16 ***
## factor(time)2003-12-31  1.058e+00  5.496e-03 192.561  < 2e-16 ***
## factor(time)2004-06-30  1.085e+00  5.468e-03 198.414  < 2e-16 ***
## factor(time)2004-12-31  1.129e+00  5.439e-03 207.574  < 2e-16 ***
## factor(time)2005-06-30  1.276e+00  5.347e-03 238.596  < 2e-16 ***
## factor(time)2005-12-31  1.367e+00  5.296e-03 258.204  < 2e-16 ***
## factor(time)2006-06-30  1.394e+00  5.282e-03 263.890  < 2e-16 ***
## factor(time)2006-12-31  1.457e+00  5.251e-03 277.440  < 2e-16 ***
## factor(time)2007-06-30  1.529e+00  5.224e-03 292.729  < 2e-16 ***
## factor(time)2007-12-31  1.541e+00  5.219e-03 295.225  < 2e-16 ***
## factor(time)2008-06-30  1.727e+00  5.140e-03 335.914  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1668357  on 532583  degrees of freedom
## Residual deviance:  409093  on 532555  degrees of freedom
## AIC: 1997733
##
## Number of Fisher Scoring iterations: 5

## Analysis of Deviance Table
##
## Model 1: Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +
##     ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta)
## Model 2: Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC + tri +
##     ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) +
##     factor(time)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    532572     825770
## 2    532555     409093 17   416678 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The associated coefficients appear to jive with expectations. There is a positive association with establishments, population, income, and metro setting. Further, tri and rural areas (non-adjacent even moreso) are associated with lower levels of broadband access. However, because the functional form of the model is non-linear I hold off on interpretation of the effects. The main point to take-away here is that the loans are not significant. To check if this is a suitable fit of a model, I turn to a histogram of fitted values and Q-Q plot for the model with time fixed effects:



Visual inspection of the fit indicates use of a count model is superior to the previous panel regression. The range of the fitted values matches that of our observed and appears to maintain a similar shape. These are far superior to any of the linear or dynamic panel methods previously used. I take this as strong evidence that a count model should be used here as opposed to a least squares panel methods.

Alternatively, I did run a regression with a time trend instead of a time fixed effect, but the model fit was worse via AIC and adjusted R-Squared.

(Quasi-)Poisson and Negative Binomial

To further inspect this relationship, two other models are considered to account for the potential overdispersion of observed variance: Quasipoisson and Negative Binomial.

```
##
## Call:
## glm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) +
##     factor(time), family = quasipoisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3988  -0.6521  -0.0568   0.4923   3.9361
##
## Coefficients:
##                           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            -2.952e+00  2.565e-02 -115.078  < 2e-16 ***
## iloans                  7.863e-03  2.957e-03    2.660  0.00782 **
## log(est)                2.503e-01  4.028e-04  621.324  < 2e-16 ***
## log(Pop_IRS)            7.518e-02  4.942e-04  152.127  < 2e-16 ***
```

```
## logINC                  1.286e-01  2.455e-03   52.366  < 2e-16 ***
## tri                    -1.249e-03  3.042e-05  -41.063  < 2e-16 ***
## rucadj                 -1.897e-02  1.781e-03  -10.652  < 2e-16 ***
## rucnonadj              -3.547e-02  2.192e-03  -16.180  < 2e-16 ***
## poly(AREA_zcta, 2)1    -3.332e+00  4.781e-01   -6.970 3.17e-12 ***
## poly(AREA_zcta, 2)2     1.390e+00  4.546e-01    3.057  0.00223 **
## I(Pop_IRS/AREA_cty)     4.201e-06  1.744e-07   24.086  < 2e-16 ***
## I(est/AREA_zcta)        1.668e-06  7.550e-07    2.209  0.02718 *
## factor(time)2000-06-30  2.185e-01  5.251e-03   41.613  < 2e-16 ***
## factor(time)2000-12-31  4.762e-01  4.979e-03   95.655  < 2e-16 ***
## factor(time)2001-06-30  6.522e-01  4.827e-03  135.119  < 2e-16 ***
## factor(time)2001-12-31  7.172e-01  4.775e-03  150.218  < 2e-16 ***
## factor(time)2002-06-30  8.397e-01  4.683e-03  179.330  < 2e-16 ***
## factor(time)2002-12-31  9.202e-01  4.628e-03  198.841  < 2e-16 ***
## factor(time)2003-06-30  1.014e+00  4.567e-03  222.053  < 2e-16 ***
## factor(time)2003-12-31  1.058e+00  4.541e-03  233.027  < 2e-16 ***
## factor(time)2004-06-30  1.085e+00  4.519e-03  240.110  < 2e-16 ***
## factor(time)2004-12-31  1.129e+00  4.494e-03  251.195  < 2e-16 ***
## factor(time)2005-06-30  1.276e+00  4.418e-03  288.735  < 2e-16 ***
## factor(time)2005-12-31  1.367e+00  4.376e-03  312.464  < 2e-16 ***
## factor(time)2006-06-30  1.394e+00  4.365e-03  319.345  < 2e-16 ***
## factor(time)2006-12-31  1.457e+00  4.339e-03  335.742  < 2e-16 ***
## factor(time)2007-06-30  1.529e+00  4.317e-03  354.244  < 2e-16 ***
## factor(time)2007-12-31  1.541e+00  4.313e-03  357.264  < 2e-16 ***
## factor(time)2008-06-30  1.727e+00  4.248e-03  406.504  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.6828514)
##
##     Null deviance: 1668357  on 532583  degrees of freedom
## Residual deviance:  409093  on 532555  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5


##
## Call:
## glm(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) + logINC +
##     tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) + I(est/AREA_zcta) +
##     factor(time), family = negative.binomial(theta = 1), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.17958  -0.28831  -0.02073   0.22569   2.23380
##
## Coefficients:
##                           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            -3.503e+00  3.354e-02 -104.458  < 2e-16 ***
## iloans                  2.206e-03  4.079e-03    0.541   0.5887
## log(est)                2.488e-01  4.619e-04  538.757  < 2e-16 ***
## log(Pop_IRS)            8.318e-02  6.400e-04  129.977  < 2e-16 ***
## logINC                  1.652e-01  3.286e-03   50.281  < 2e-16 ***
## tri                    -7.756e-04  3.430e-05  -22.612  < 2e-16 ***
```

```
## rucadj                 -1.736e-02  2.017e-03   -8.607  < 2e-16 ***
## rucnonadj              -3.777e-02  2.416e-03  -15.634  < 2e-16 ***
## poly(AREA_zcta, 2)1     8.377e-01  5.229e-01    1.602   0.1092
## poly(AREA_zcta, 2)2    -1.295e+00  5.079e-01   -2.549   0.0108 *
## I(Pop_IRS/AREA_cty)     9.792e-06  3.065e-07   31.949  < 2e-16 ***
## I(est/AREA_zcta)        7.020e-06  1.420e-06    4.942 7.74e-07 ***
## factor(time)2000-06-30  2.281e-01  4.528e-03   50.383  < 2e-16 ***
## factor(time)2000-12-31  4.703e-01  4.439e-03  105.952  < 2e-16 ***
## factor(time)2001-06-30  6.572e-01  4.384e-03  149.894  < 2e-16 ***
## factor(time)2001-12-31  7.349e-01  4.363e-03  168.448  < 2e-16 ***
## factor(time)2002-06-30  8.533e-01  4.335e-03  196.822  < 2e-16 ***
## factor(time)2002-12-31  9.396e-01  4.315e-03  217.745  < 2e-16 ***
## factor(time)2003-06-30  1.036e+00  4.294e-03  241.263  < 2e-16 ***
## factor(time)2003-12-31  1.087e+00  4.284e-03  253.614  < 2e-16 ***
## factor(time)2004-06-30  1.120e+00  4.273e-03  262.052  < 2e-16 ***
## factor(time)2004-12-31  1.166e+00  4.265e-03  273.316  < 2e-16 ***
## factor(time)2005-06-30  1.303e+00  4.242e-03  307.239  < 2e-16 ***
## factor(time)2005-12-31  1.418e+00  4.224e-03  335.577  < 2e-16 ***
## factor(time)2006-06-30  1.499e+00  4.216e-03  355.537  < 2e-16 ***
## factor(time)2006-12-31  1.592e+00  4.204e-03  378.656  < 2e-16 ***
## factor(time)2007-06-30  1.693e+00  4.206e-03  402.397  < 2e-16 ***
## factor(time)2007-12-31  1.695e+00  4.206e-03  403.070  < 2e-16 ***
## factor(time)2008-06-30  1.882e+00  4.177e-03  450.597  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.1712192)
##
##     Null deviance: 368453  on 532583  degrees of freedom
## Residual deviance: 129926  on 532555  degrees of freedom
## AIC: 2579674
##
## Number of Fisher Scoring iterations: 5


##
## Call:
## glm.nb(formula = Prov_num ~ iloans + log(est) + log(Pop_IRS) +
##     logINC + tri + ruc + poly(AREA_zcta, 2) + I(Pop_IRS/AREA_cty) +
##     I(est/AREA_zcta) + factor(time), data = data, init.theta = 66197.57996,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3987  -0.6520  -0.0568   0.4923   3.9360
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.952e+00  3.104e-02 -95.091  < 2e-16 ***
## iloans                  7.864e-03  3.578e-03   2.198   0.0280 *
## log(est)                2.503e-01  4.875e-04 513.407  < 2e-16 ***
## log(Pop_IRS)            7.518e-02  5.981e-04 125.705  < 2e-16 ***
## logINC                  1.286e-01  2.971e-03  43.271  < 2e-16 ***
## tri                    -1.249e-03  3.681e-05 -33.931  < 2e-16 ***
## rucadj                 -1.897e-02  2.155e-03  -8.801  < 2e-16 ***
```

```
## rucnonadj              -3.547e-02  2.653e-03 -13.369  < 2e-16 ***
## poly(AREA_zcta, 2)1    -3.332e+00  5.786e-01  -5.759 8.47e-09 ***
## poly(AREA_zcta, 2)2     1.390e+00  5.502e-01   2.526   0.0115 *
## I(Pop_IRS/AREA_cty)     4.202e-06  2.111e-07  19.904  < 2e-16 ***
## I(est/AREA_zcta)        1.669e-06  9.137e-07   1.826   0.0678 .
## factor(time)2000-06-30  2.185e-01  6.354e-03  34.386  < 2e-16 ***
## factor(time)2000-12-31  4.762e-01  6.025e-03  79.043  < 2e-16 ***
## factor(time)2001-06-30  6.522e-01  5.841e-03 111.652  < 2e-16 ***
## factor(time)2001-12-31  7.172e-01  5.778e-03 124.130  < 2e-16 ***
## factor(time)2002-06-30  8.397e-01  5.667e-03 148.185  < 2e-16 ***
## factor(time)2002-12-31  9.202e-01  5.601e-03 164.308  < 2e-16 ***
## factor(time)2003-06-30  1.014e+00  5.527e-03 183.488  < 2e-16 ***
## factor(time)2003-12-31  1.058e+00  5.496e-03 192.556  < 2e-16 ***
## factor(time)2004-06-30  1.085e+00  5.468e-03 198.409  < 2e-16 ***
## factor(time)2004-12-31  1.129e+00  5.439e-03 207.569  < 2e-16 ***
## factor(time)2005-06-30  1.276e+00  5.347e-03 238.589  < 2e-16 ***
## factor(time)2005-12-31  1.367e+00  5.296e-03 258.197  < 2e-16 ***
## factor(time)2006-06-30  1.394e+00  5.282e-03 263.885  < 2e-16 ***
## factor(time)2006-12-31  1.457e+00  5.251e-03 277.435  < 2e-16 ***
## factor(time)2007-06-30  1.529e+00  5.224e-03 292.724  < 2e-16 ***
## factor(time)2007-12-31  1.541e+00  5.219e-03 295.220  < 2e-16 ***
## factor(time)2008-06-30  1.727e+00  5.140e-03 335.908  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(66197.53) family taken to be 1)
##
##     Null deviance: 1668235  on 532583  degrees of freedom
## Residual deviance:  409066  on 532555  degrees of freedom
## AIC: 1997746
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  66198
##           Std. Err.:  10932
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -1997686
```

There is some problem that I cannot figure out with the Negative Binomial model as it gives identical results to the Poisson Model. Effectively, the Negative Binomial model is fit via maximum likelihood methods and the numerical algorithm does not converge. I do not know what is causing this, although the Negative Binomial model will converge if I remove time fixed effects. I err on the side of avoiding the Negative Binomial as opposed to dropping the time fixed effects.

As for the Quasi-Poisson Model, the only difference appears to be that the standard errors are smaller. The coefficients across all models are identical, which also results in residuals being identical across all models.

Residuals

We can turn to inspection of time, space, and space-time correlation in residuals to determine the fit of our model and that assumptions are satisfied. First, the correlation structure of residuals across time:

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1999	1.000	0.676	0.435	0.358	0.289	0.206	0.176	-0.003	-0.091	-0.110
2000	0.676	1.000	0.689	0.520	0.407	0.299	0.198	-0.035	-0.092	-0.106
2001	0.435	0.689	1.000	0.764	0.577	0.412	0.252	0.014	-0.043	-0.043
2002	0.358	0.520	0.764	1.000	0.789	0.582	0.359	0.076	0.014	0.025
2003	0.289	0.407	0.577	0.789	1.000	0.759	0.450	0.150	0.073	0.082
2004	0.206	0.299	0.412	0.582	0.759	1.000	0.590	0.255	0.162	0.158
2005	0.176	0.198	0.252	0.359	0.450	0.590	1.000	0.517	0.343	0.323
2006	-0.003	-0.035	0.014	0.076	0.150	0.255	0.517	1.000	0.782	0.648
2007	-0.091	-0.092	-0.043	0.014	0.073	0.162	0.343	0.782	1.000	0.836
2008	-0.110	-0.106	-0.043	0.025	0.082	0.158	0.323	0.648	0.836	1.000

(I have averaged the residuals by year across time so that the correlation structure can be visible.)

One can observe that there is a pattern where the correlation across all observations each year have a higher correlation to the closer years than further away. So there is evidence of serial correlation, although it is unclear how to proceed because this is a relatively short panel compared to the number of observations. Fitting an autoregressive (AR) or moving average (MA) process for each zip code would introduce a large number of parameters and I begin to wonder if the cure is worse than the disease in this scenario.

The typical applied remedy is to simply use robust standard errors in significance testing. The `vcovHAC` function in `R` allows for heteroskedastic and autocorrelation consistent covariance matrix estimation. Applying this to our previous models would yield different standard errors, but the point estimates would remain the same. Seeing that the Poisson, Quasi-Poisson, and Negative Binomial resulted in the same coefficients, we can use the `vccovHAC` function on the Poisson regression to observe this effect. Unfortunately, as I have attempted to run this in `R` I have not been able to get the function to work. The model must be too large for my computer to handle, so instead I will present the `vcovHC` results though which are the White standard errors:

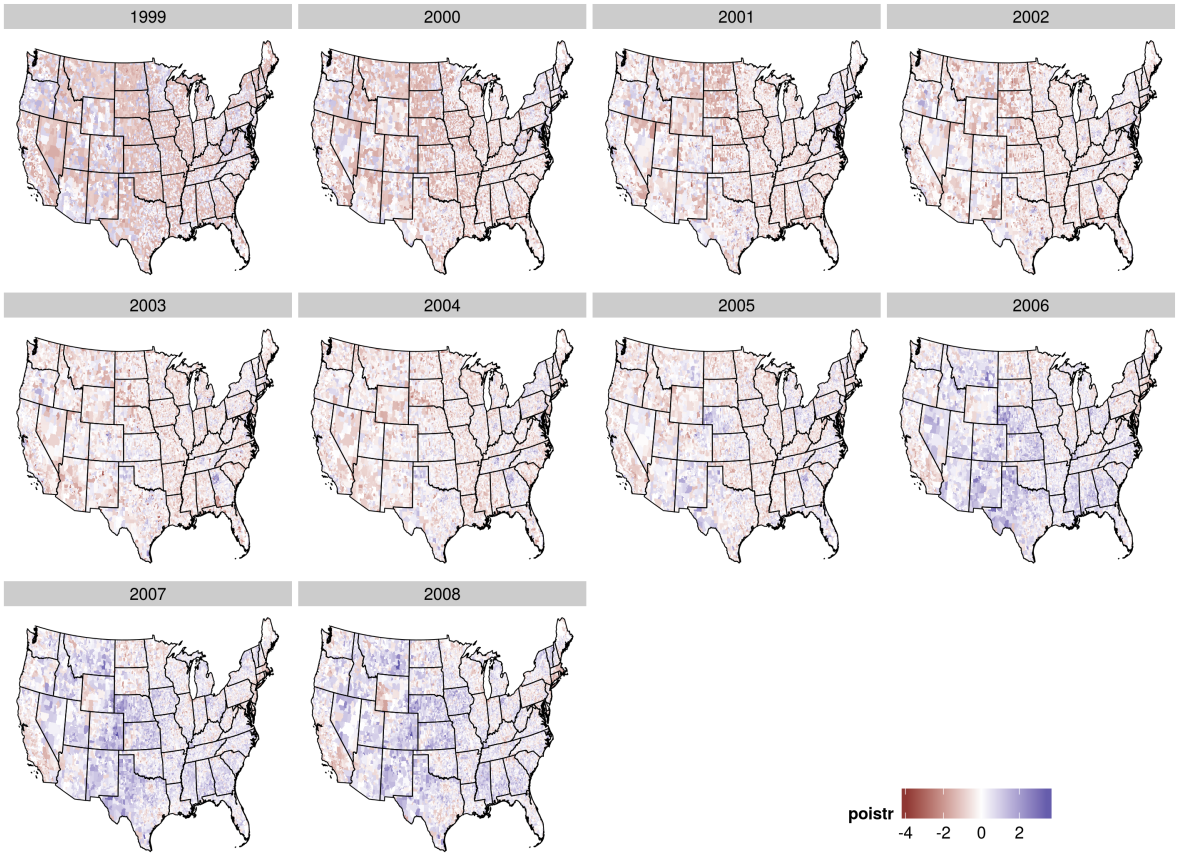
```
##
## z test of coefficients:
##
##                            Estimate  Std. Error   z value  Pr(>|z|)
## (Intercept)            -2.9517e+00  2.7435e-02 -107.5890 < 2.2e-16 ***
## iloans                  7.8631e-03  2.8290e-03    2.7794  0.005445 **
## log(est)                2.5029e-01  4.0288e-04  621.2606 < 2.2e-16 ***
## log(Pop_IRS)            7.5180e-02  5.4299e-04  138.4569 < 2.2e-16 ***
## logINC                  1.2857e-01  2.6658e-03   48.2307 < 2.2e-16 ***
## tri                    -1.2491e-03  2.9110e-05  -42.9120 < 2.2e-16 ***
## rucadj                 -1.8968e-02  1.6585e-03  -11.4363 < 2.2e-16 ***
## rucnonadj              -3.5474e-02  2.1040e-03  -16.8604 < 2.2e-16 ***
## poly(AREA_zcta, 2)1    -3.3324e+00  4.5985e-01   -7.2467 4.270e-13 ***
## poly(AREA_zcta, 2)2     1.3900e+00  3.5497e-01    3.9159 9.007e-05 ***
## I(Pop_IRS/AREA_cty)     4.2011e-06  2.2011e-07   19.0866 < 2.2e-16 ***
## I(est/AREA_zcta)        1.6676e-06  1.0023e-06    1.6637  0.096173 .
## factor(time)2000-06-30  2.1849e-01  5.5555e-03   39.3286 < 2.2e-16 ***
## factor(time)2000-12-31  4.7623e-01  5.5117e-03   86.4037 < 2.2e-16 ***
## factor(time)2001-06-30  6.5220e-01  5.2272e-03  124.7692 < 2.2e-16 ***
## factor(time)2001-12-31  7.1723e-01  5.0665e-03  141.5629 < 2.2e-16 ***
## factor(time)2002-06-30  8.3973e-01  5.0159e-03  167.4143 < 2.2e-16 ***
## factor(time)2002-12-31  9.2022e-01  4.9381e-03  186.3512 < 2.2e-16 ***
## factor(time)2003-06-30  1.0142e+00  4.8489e-03  209.1574 < 2.2e-16 ***
## factor(time)2003-12-31  1.0582e+00  4.8105e-03  219.9824 < 2.2e-16 ***
```

```
## factor(time)2004-06-30  1.0850e+00  4.6989e-03  230.9049 < 2.2e-16 ***
## factor(time)2004-12-31  1.1290e+00  4.6681e-03  241.8546 < 2.2e-16 ***
## factor(time)2005-06-30  1.2757e+00  4.5985e-03  277.4064 < 2.2e-16 ***
## factor(time)2005-12-31  1.3675e+00  4.5563e-03  300.1262 < 2.2e-16 ***
## factor(time)2006-06-30  1.3938e+00  4.5010e-03  309.6742 < 2.2e-16 ***
## factor(time)2006-12-31  1.4567e+00  4.5742e-03  318.4633 < 2.2e-16 ***
## factor(time)2007-06-30  1.5293e+00  4.5699e-03  334.6460 < 2.2e-16 ***
## factor(time)2007-12-31  1.5408e+00  4.5362e-03  339.6741 < 2.2e-16 ***
## factor(time)2008-06-30  1.7266e+00  4.4808e-03  385.3399 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is slightly puzzling that the White standard errors are generally smaller than the other models. This gives me pause as to how to approach the calculation of standard errors in this setting ... but onto the spatial effects.

Spatial Effects

As I have mentioned before, the diffusion process of broadband is decidedly spatial from an engineering perspective. For an area to receive broadband access, various hubs-cables-lines need to be built. This is not unlike electrification, telephones, and highways. So as a quick check to see if this is an actual problem would be through some maps:

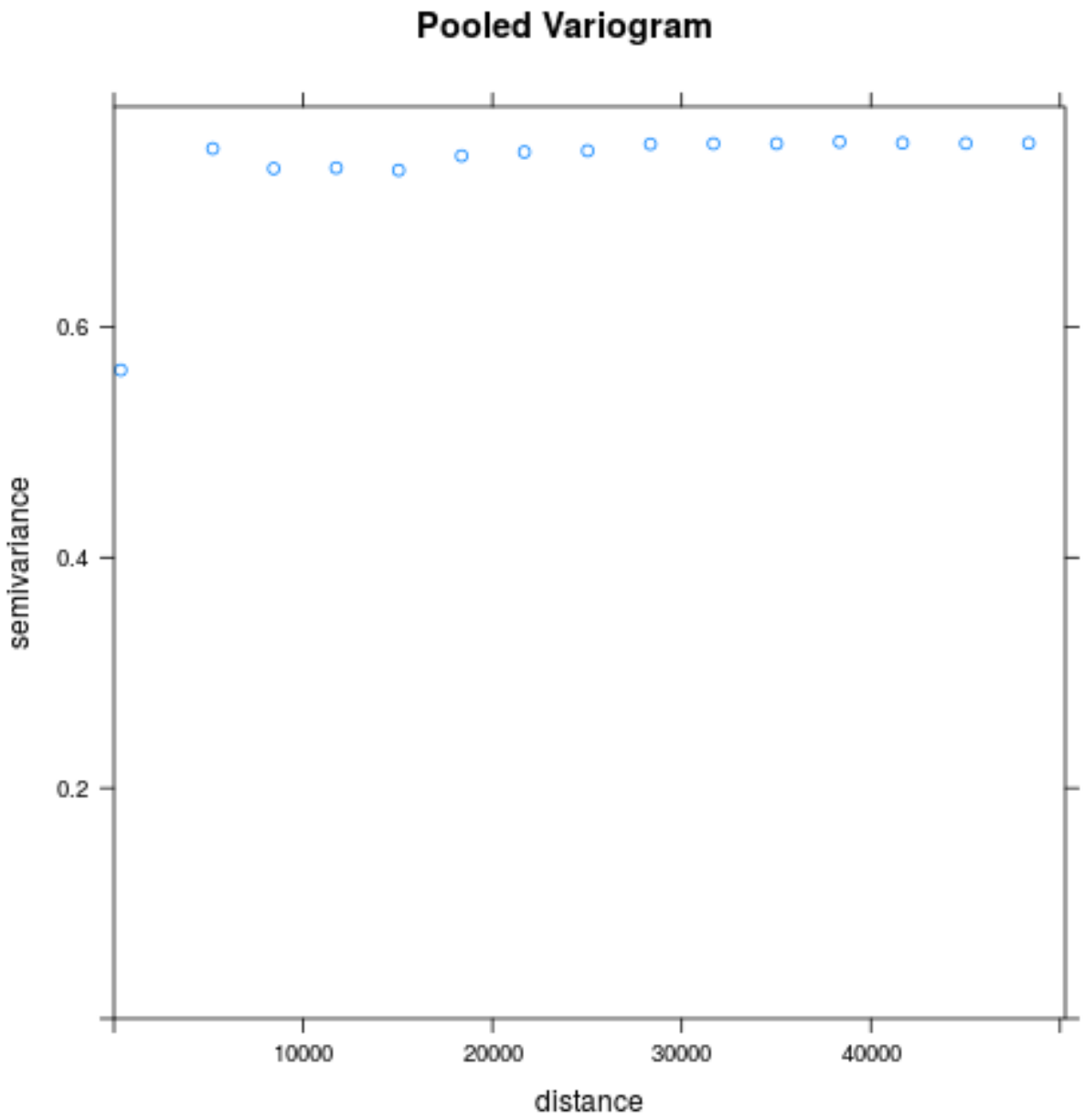


Does there appear to be spatial clustering of residuals? Well, it is possible but my eyesight is not the best. A more sophisticated way to evaluate the potential for spatial autocorrelation is through the use of an empircal variogram. A variogram is defined as the variance of the difference between field values at two locations across realizations of the field (Cressie 1993). There are theoretical variograms that describe a spatial process via parameters which can be estimated, however before getting to that step it is best to look at the empirical variogram which is defined as:

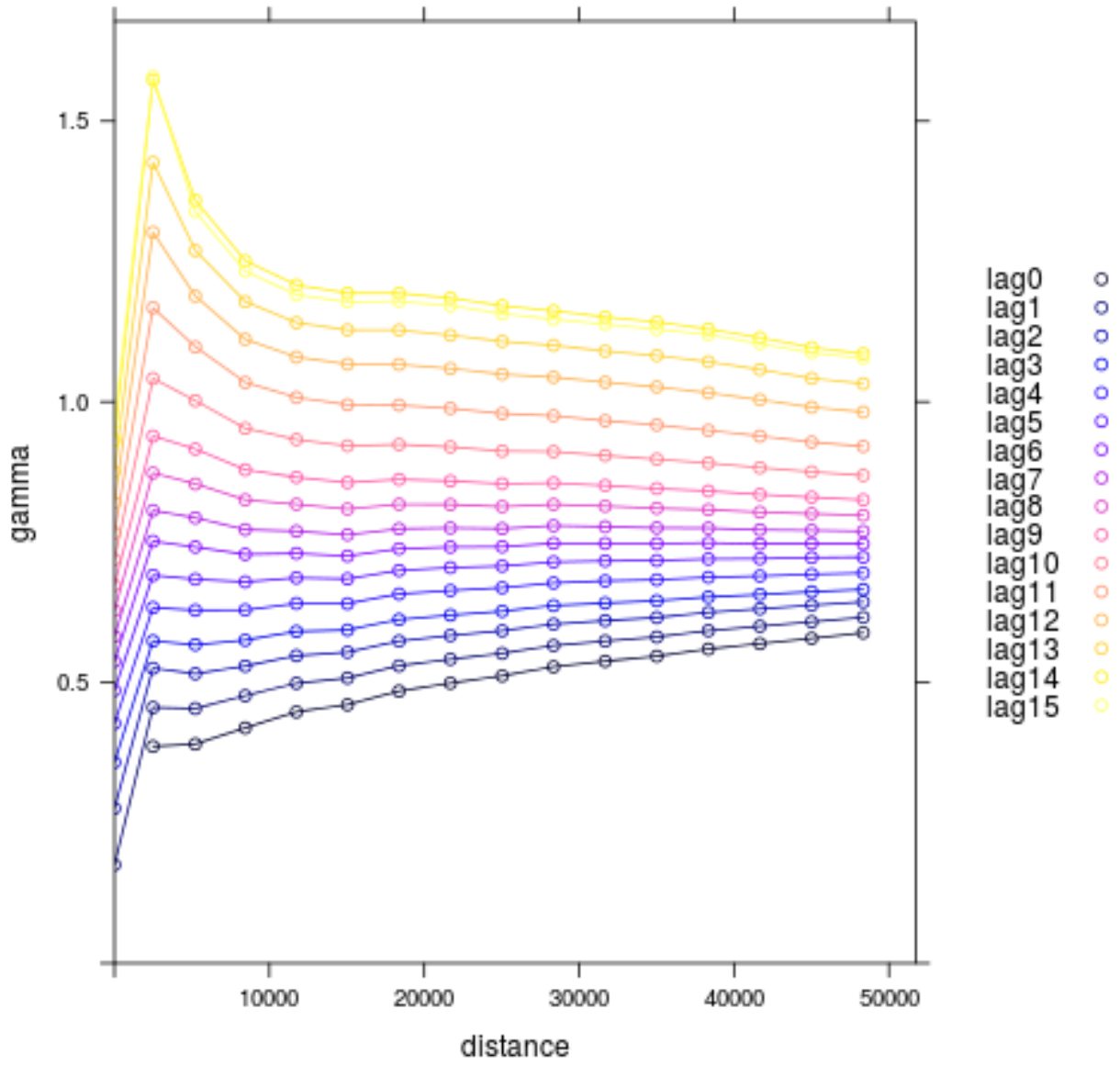
$$\gamma(h) = Var(Y(s+h) - Y(s))$$

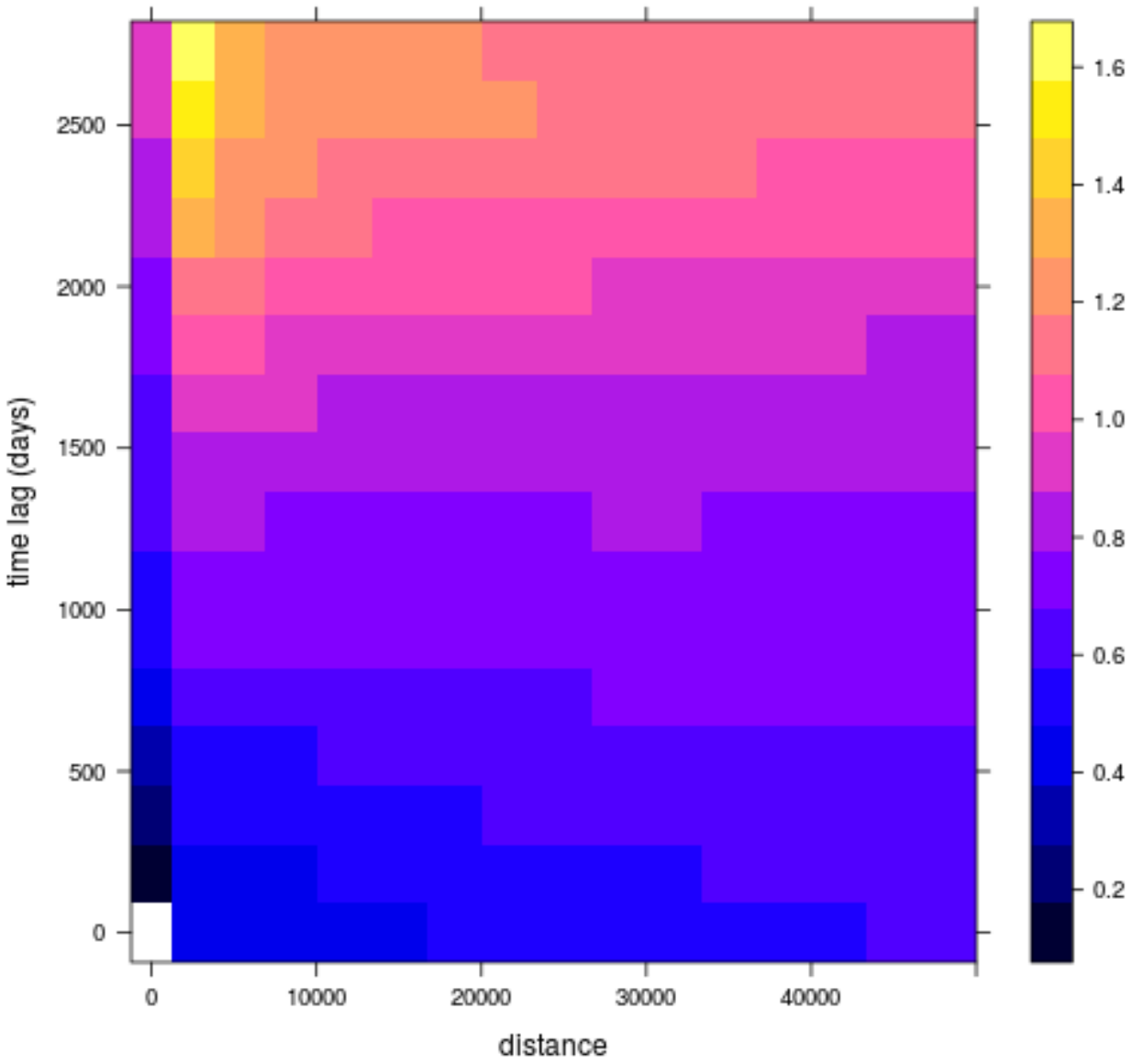
where s denotes the location of a spatial process (usually with longitude and latitude) and h is the difference between two locations. This is the analogue to a partial autocorrelation function in the time series methods, except a variogram is the inverse of the partial autocorrelation. A typical empirical variogram of a spatial process will start with a low value on the variogram (indicating high correlation) and then increase up to a particular value given a distance and then appear to be flat from there on. If this happens, then it indicates that closer values are closely related while further away values are not as related.

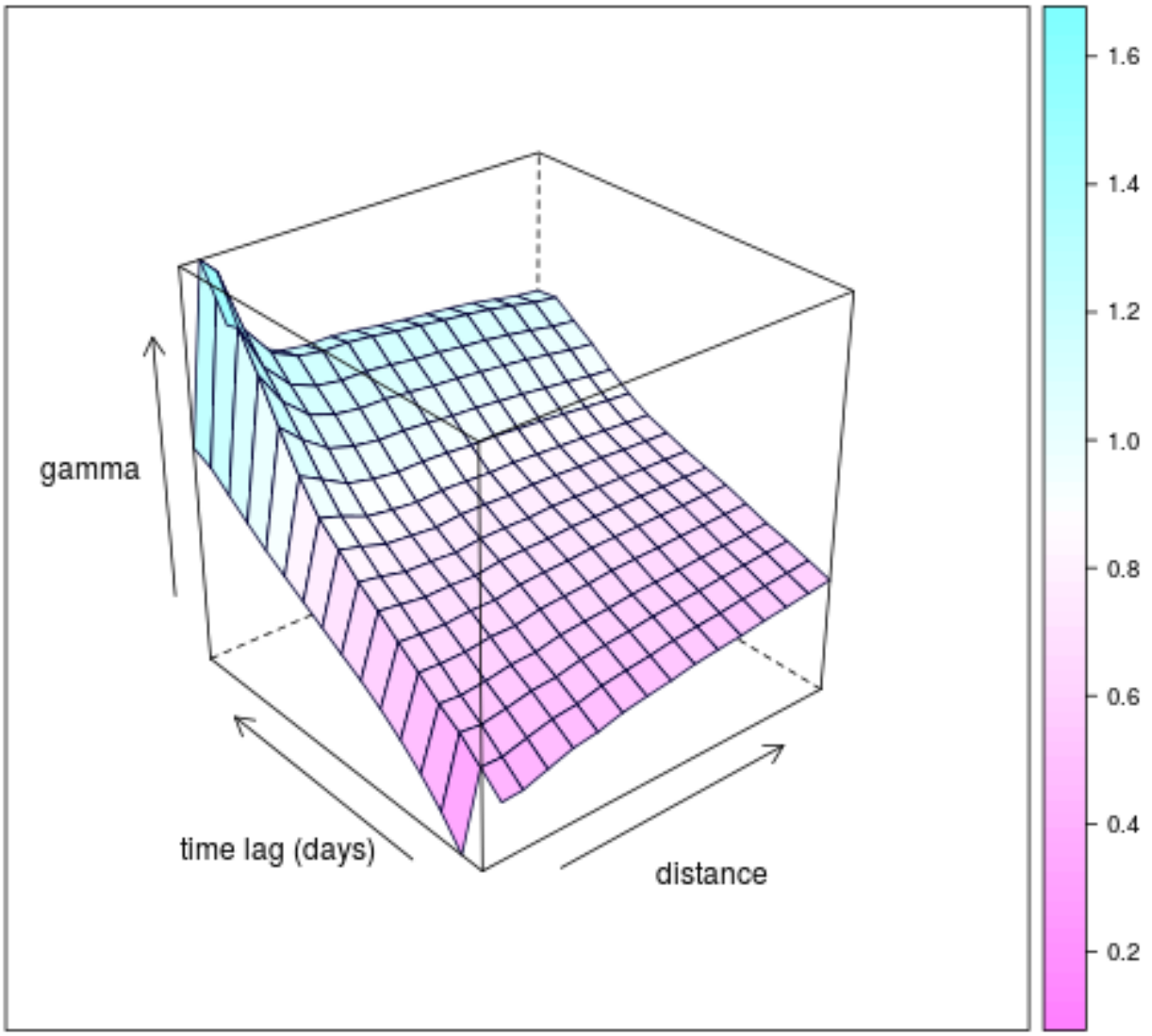
Here is the pooled empirical variogram where time is ignored and the residuals of the Poisson model are clumped together and then the distance between are binned up and plotted:



Huh. Well the distance here is in terms of meters and it doesn't appear that there is any pattern in the residuals due to distance from others. The semivariance appears to be fairly constant across all distances, which would make fitting a spatial model pointless. But it might be the case that the temporal autocorrelation is masking the spatial effects. In order to evalute this, we can compile an empirical spatio-temporal variogram. A three dimensional variogram so to speak:


gamma
1.5
1.0
0.5
10000
20000
30000
40000
50000
distance
lag0
lag1
lag2
lag3
lag4
lag5
lag6
lag7
lag8
lag9
lag10
lag11
lag12
lag13
lag14
lag15


time lag (days)
2500
2000
1500
1000
500
0
0
10000
20000
30000
40000
distance
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2



OK at this point it seems like going down the path of modeling the diffusion of broadband providers via spatio-temporal methods is one that does not bear much benefit. This may also provide insight into why I have not been able to estimate any meaningful spatial model to this data.