TI 2000-036/3
Tinbergen Institute Discussion Paper

An Empirical Measure for Labor Market Density

Pieter A. Gautier
Coen N. Teulings

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Most TI discussion papers can be downloaded at
http://www.tinbergen.nl

An empirical measure for labor market density*

Pieter A. Gautier†
Coen N. Teulings ‡

April 12, 2000

Abstract

In this paper we derive a structural measure for labor market density based on the Ellison and Glasear (1997) index for industry concentration". This labor market density measure serves as a proxy for the number of workers that can reach a certain work area within a reasonal amount of traveling time. We apply this measure to a standard wage equation and find that it takes account of almost half of the cross region wage variance (not explained by other observables). Moreover, it explains substantiallly more than the traditional density measure: people per square mile.

Keywords: labor market density, wage equation
JEL codes: J210, J300, J600, J230

*Most of this research has been carried out at the Industrial relations office in Princeton. We thank Alan Krueger and seminar participants at the Tinbergen Institute for useful comments and David Jaeger for kindly providing us with his program to map county groups to (C)MSA's.
†Erasmus University Rotterdam and Tinbergen Institute Amsterdam; gautier@tinbinst.nl
‡Erasmus University Rotterdam and Tinbergen Institute; teulings@tinbinst.nl

1 Introduction

Search frictions play an important role in the labour market. Job seekers and vacancies do not meet instantaneously, their matching takes effort and time. However, how much time search takes depends on the characteristics of the labor market. An obvious factor that matters is the density of a labor market: the more job seekers and vacancies are available in an area, the easier it is for them to find an acceptable match. Several authors have developed models along these lines, see e.g.: Diamond (1982), Burda and Profit (1996), Coles and Smith (1994,98) and Wasmer and Zenou (1999). Although there is a large literature that suggests that returns to scale in job search are constant, there are at least three reasons why the number of job seekers and vacancies might matter. First, workers who live in an area with many vacancies have a larger set of feasible jobs to choose from. We expect that there will be fewer mismatches and shorter unemployment spells after displacement in those areas. Second, workers with a larger set of feasible jobs have more bargaining power and are therefore likely to earn higher wages. Third, if workers are mobile, arbitrage will equalize reservation wages within skill groups of workers across regions. This implies that the worker and job types who gain the most from low search costs move to areas where the contact rate is high. In Teulings and Gautier (2000), we argue that those are typically the workers with the highest and the lowest skills because the market is relatively thin for them.

A big obstacle in research in this area is that labor market density is difficult to measure. One likely candidate is simply the amount of workers and/or jobs per square mile. However, a number of serious drawbacks to this measure immediately come to mind. First, it ignores the role of infrastructure. What we are really interested in is not the set of applicants within a certain distance of the job, but within, say one hour commuting time. The relevant labor market area should then be weighted by the number of highways and public transport facilities. Moreover, distance is not the only factor. When particular locations are more attractive for living while others have an advantage as work area, people might be prepared to accept on average a longer commuting time.

These considerations suggest that we should look for a measure based on revealed preferences. The measure that we propose is not based on weighted commuting distance or time, but on commuting patterns that we can actually observe. The idea is that we take the location of the job of a worker as given and then analyze where the worker decides to live. If we observe that all workers live in the same area as where they work, a given job can only be occupied by a limited number of workers. This is typical for a small scale labor marked. Alternatively, when workers working in a particular location live in many different areas, the scale of the labor market is large. More specifically, our measure can be viewed of as a model based index of geographic labor market density (or reachability) similar to the dartboard index for industry concentration of Ellison and Glasear

(1997, EG from now onwards). The index can take any value between zero and one. When it is equal to one, the labor market is hard to reach and the only workers who work in a particular area are the ones who live there. When it is equal to zero, the labor market is extremely easy to reach and we observe workers from many different areas to be employed in this labor market. The measure has the advantage that it controls for the size of the area on which it is defined so that one can meaningfully compare results from different data sets (with different levels of aggregation) with each other.

The plan of the paper is as follows. Section 2 derives the index from location decisions of utility maximizing agents. Section 3 describes how the index can be constructed from the 5% public use micro samples of the Census and how it can be linked to the CPS. Finally, section 4 gives an illustration in the form of a wage equation. It is a well known fact that there exists substantial cross-regional variation in wages. We find that almost 50% of the regional variation is captured by our density measure. Moreover, we find that our measure does a substantially better job in explaining this variation than the number of persons per square mile.

2 The index

Consider the decision problem for the k th worker with a job in area w who has to choose an area v_k to live in. Let the utility for area h be given by:

$$\log \pi_{kwh} = \log \bar{\pi}_{wh} + \varepsilon_{kwh} \tag{1}$$

where the ε_{kwh}'s reflect idiosyncratic factors (like the relative preference for clean air, safety, theater availability etcetera) which are assumed to be independent Weibull random variables which are independent of $\{\bar{\pi}_{wh}\}$, and $\bar{\pi}_{wh}$ is a random location specific variable, which is chosen by nature at the start of the process. It reflects the attractiveness to live in a certain area (given that the agent's job is in w) for a typical agent. Conditional on the realization of $\bar{\pi}_{w1}, ..., \bar{\pi}_{wH}$and given our assumptions on ε_{kh} we can write:

$$Prob\{v_k = h| \, \bar{\pi}_{wh}, ..., \bar{\pi}_{wH}\} = \frac{\bar{\pi}_{wh}}{\sum_j \bar{\pi}_{wj}}$$

which is a conditional logit model, see McFadden (1973). Next, we make the same parametric restrictions on the distribution of the $\{\bar{\pi}_{wh}\}$ as EG. First, we want that on average the model reproduces the overall distribution of residence (i.e., it puts more workers in New York than in a small village). Therefore, assume that:

$$E_{\pi_{w1}, ..., \pi_{wM}} \frac{\bar{\pi}_{wh}}{\sum_j \bar{\pi}_{wj}} = x_h \tag{2}$$

where, x_h, is the relative size of area h (fraction of total population who lives in h). Second, we have to make assumptions with regard to the relative importance of "reachability" to the agents. Let the joint distribution of $\bar{\pi}_{wh}$ be such that there is a single parameter $\gamma_w \in [0,1]$ for which

$$var\left(\frac{\bar{\pi}_{wh}}{\sum_j \bar{\pi}_{wj}}\right) \equiv v_w = \gamma_w x_h(1 - x_h) \tag{3}$$

The variance v_w measures how sensitive the agent's utility is to a good fit. For jobs in rural areas, the variance is likely to be high because those jobs are typically hard to reach and therefore the utility of living in another area than the area where one's job is located will be small. So the few areas that are within reasonably traveling distance from the work area have high $\bar{\pi}_{wh}$'s, the rest of the areas have $\bar{\pi}_{wh}$= 0. When $\gamma_w = 1$, the variance, v_w, reaches a maximum (since the maximum variance of a variable with mean x_h that lies between zero and one is $x_h\,[1 - x_h]$). The variation in idiosyncratic characteristics ε_{kwh} is dominated by the variation in the location specific factors, $\log \bar{\pi}_{wh}$. When $\gamma_w = 0$, the location decision is totally dominated by the agent's idiosyncratic taste factors. The agent's decision on where to live is independent of the location of the job and each living area h is chosen with probability x_h. The parameter γ_w therefore captures the importance of regional factors relative to idiosyncratic taste factors of the agents.

Now we will define an unbiased estimator for γ_w. Let s_{wh} be the number of workers working in area w and living in area h as a share of the total employment in area w. The following relation applies between γ_w on the one hand and s_{wh} and the sizes of the areas of residence x_h on the other hand.

Proposition 1 *In any specification of the location choice model in which agents 1,2,...,N choose locations to maximize utility that satisfy equations (2), and (3), an unbiased estimator for γ_w is:*

$$\gamma_w = \frac{\sum_h (s_{wh} - x_h)^2}{(1 - \sum x_h^2)} \tag{4}$$

Proof: *See appendix 1.*

This proposition is a special case of EG's Proposition 1. To illustrate how this measure is related to the scale of the labor market, consider a job in area w. Let N be the total population and let n be the number of workers who is willing to work in area w and let all of them have an equal probability to get this job. Hence, n is a measure for the scale of the labor market. Their probability to get this job is $1/n$ and the probability for the rest of the population, $N - n$, to get the job is equal to zero. Hence, a fraction $(1 - n/N)$ of the population has a zero probability to work in w and a fraction n/N has a probability $1/n$. Since the variance of the binomial distribution for a stochast taking the values $(0, b)$ is $b^2 p(1 - p)$, the variance of this process is: $V = (1/n)^2[(1-n/N)n/N] = 1/N[1/n - 1/N]$. Since $V = \gamma\frac{1}{N}(1-\frac{1}{N})$, we get for $N \to \infty$, $\gamma \simeq \frac{1}{n}$. Hence, in this simple binomial example where workers either do or do not belong to a market and where all workers in a market have an equal probability for a particular job, γ is equal to the inverse of the scale of the labor market.

The above analysis takes as a starting point the work area of the worker and then analyses the choice of the optimal living area. We could also have proceeded the other way around, by analyzing the choice of the optimal work area conditional on the living area. Our actual conditioning on work area in our calculations is based on the idea that work areas can be heavily concentrated in city centres. Then, conditioning on living area would underestimate the density of the city centres. Most people living in Manhattan are likely to work in Manhattan, incorrectly suggesting that Manhattan is a low density area. However, most people working in Manhattan live in other regions. Hence, by conditioning on work areas we avoid the problem of the mismeasurement of γ_w in city centres.

An advantage of this measure is that it is easy to calculate. All one needs is data with information for a set of workers on the location of their job and their home. We do not need to know the spatial relations between all regions, all this information is embedded in the data. However, there is one problem. Ideally, this measure is independent of the level of aggregation of the location measure. Whether one measures location at the state level or county level should not affect the calculated value of γ_w for a state. However, this requires that the values of $\{\bar{\pi}_{wh}\}$ are drawn independently of the aggregation scheme of subregions into regions. Obviously, this assumption is violated in our application. Any aggregation will merge adjacent sub-regions into a new region. Hence, the values of $\{\bar{\pi}_{wh}\}$ for sub-regions within a region will be highly correlated. The consequences of this can be seen easily by considering the limiting effect of the aggregation of all subregions into a single region. All workers will live in the area where they work and hence γ_w will be equal to unity. In general, aggregation will therefore tend to reduce the estimate of γ_w. As long as the number of regions is large and the sizes of the regions do not vary too much, this problem is not likely to greatly affect the relative sizes of the calculated γ_w's. In the next section, we present calculations of γ_w from Census data. Aggregation bias of the sort described above does not seem to play an important role since for this particular application we did not find γ_w to be higher in large areas.

3 Data

3.1 Constructing the index from census data

The US Census data are well suited for the construction of our measure because they contain detailed information on both the area of residence and the work area at low levels of aggregation. We use the 5% public use micro samples (PUMS) of the 1990 census. The most disaggregate geographic unit in the census is the Public Use Micro data Area (PUMA). A typical PUMA is populated by at least 100,000 persons and is identified by a five-digit number which is unique within states. In dense areas, PUMA's define a subset of a single county while in the rural states, PUMA's consist of a number of different counties. To construct our density measure we also need information on the area where the worker works (PUMAW). This is however defined at the 2 digit level, which corresponds exactly to the first 2 digits of the PUMA's of residence. The analysis will therefore be on 2-digit PUMA's With the method of the previous section we were able to construct a γ_w for each of the 1138 2-digit PUMA's.[1]

In calculating γ_w, we included only the workers who were full time employed in the US and who did not live or work in Alaska or Hawaii.. Since in general, each area is very small compared to the whole country, the denominator of (4) is close to one (i.e. using Census data, we found for the US: $\sum_w x_h^2 = 0.0024$) and γ_w is therefore almost entirely determined by $\Sigma_h \left(s_{wh} - x_h\right)^2$. To get an idea of the range of possible values γ_w can obtain, we found γ_w to be equal to 0.07 in Northern New Jersey while for some areas in Arizona, Maine, Missouri, Montana, Kansas and Wyoming we found values of γ_w as high as 0.95.

A simple OLS regression of (log) γ_w on the (log) relative size of the area shows that there exists a negative relation between γ_w and relative area size (the elasticity = -0.1, s.d.=0.02). When aggregation bias would have been important, this relation should be positive (see the discussion in the previous section). The reason for the negative relation is most likely that central city areas are both larger (in terms of inhabitants) and easier to reach than non-central city areas.

Finally, since $\{\bar{\pi}_{wh}\}$ are not independent we do not want the standard deviation of the size of the PUMA's to be too large. This is luckily not the case. Both the mean and the standard deviation of the relative PUMA size are 0.001.

Figure 1 plots the size distribution.

FIGURE 1 ABOUT HERE

3.2 Using additional information from the CPS

For many economic applications, the CPS contains crucial individual information which is not present in the Census. The CPS does however not contain information on the work location. We therefore link the Census based $\gamma_w's$ to the place of residence in the CPS. This is not a trivial operation because there is no one to one match between the PUMA's (public use micro area) of the census and the CMSA/M(S)A (central metropolitan area) and state classification of the CPS. We therefore use the following strategy to map the PUMAW to the (C)MSA 's of the CPS. First, we match the PUMAW's to MSA/CMSA 's, using the method

[1]We restricted our analysis to the workers who were employed.

of Jaeger et al. (1997). We aggregate by taking weighted (by relative area size) averages of the relevant γ_w's.[2] In most states there are however areas which do not belong to a CMSA/MSA. Those are typically rural areas. For those areas we also calculated weighted average γ_w $'s$ per state.[3] This leaves us with in total 182 unique γ_w's. To illustrate this aggregation procedure, consider the following example for Indianapolis, IN. At the 2-digit PUMA level, the Indianapolis CMSA, consists of four PUMA's, each with a unique γ_{Census}. In the CPS, Indianapolis is treated as a single geographical unit. We take weighted (by x_w) averages of γ_{Census} to get a unique γ_{CPS} for Indianapolis.

PUMA	CMSA, state	γ_{Census}	x_w	weight	γ_{CPS}
1	Indianapolis, IN	0.54221	0.004411	0.76854	0.53501
33	Indianapolis, IN	0.53478	0.000289	0.05029	0.53501
34	Indianapolis, IN	0.56212	0.000387	0.06737	0.53501
35	Indianapolis, IN	0.47045	0.000653	0.11380	0.53501

Thus, although the geographical measures of the CPS are less detailed than the ones of the census, we do use the disaggregate information as much as possible. Figure 1 depicts the density of γ_w for the 1138 Census areas while Figure 2 plots γ_w for the 182 CPS areas. The mean for the Census γ_w is 0.597 and the standard deviation is 0.235 while for the CPS those values are respectively 0.586 and 0.217. Whereas the weighted (by area size) mean for the Census γ_w is 0.539 while it is equal to 0.540 for the CPS γ_w. Hence, we do not loose much variation in our measure by this spatial aggregation.

Figure 1 about here

Figure 2 about here

We expect γ_w to be related to population density (measured in persons per square mile) and the amount of highways and railroads in an area. Figures 3 and 4 are illustrative in this respect. Figure 3 shows a map of all the counties in the U.S., where the darker areas are more densely populated. In this Figure we inserted some values of γ_w, based on the Census public use micro areas. We clearly see that densely populated areas have smaller $\gamma_w' s$.[4] The correlation between γ_w and the amount of people per square mile is -0.43. If we compare the

[2]We made some slight adjustments in their program since we observe only 2 digit PUMAW's and the CMSA/MSA's of the CPS and Census do not match exactly. For example, in the CPS, Denver and Houston have respectively the numbers: 2080 and 3060, while in the census those numbers are 2082 and 3062. For most cases, changing the last digit into zero was sufficient, only for Miami , the CMSA is 5000 in the CPS and 4992 in the Census.

[3]For the definitions of (C)M(S)A's we refer to the apendix. Our density measures and relevant weights per PUMAW of the 1990 census and per (C)MSA/MA of the CPS, and SAS formats for (C)MSA's and states can be found at: http: // www2.tinbergen.nl / ˜gautier / lm-density.html. We present aggregation results for complete (C)M(S)A's and for (C)M(S)A*state area's. In the first case, Northern New Jersey is included in the NY-CMSA, whereas in the second case it is not.

[4]This picture is mainly illustrative for the relation between γ_w and population density because the larger cities sometimes consist of multiple counties and PUMA's.

cities from the East Coast with those from the West Coast, we see that jobs in the more densely populated East Coast cities are easier to reach since γ_w tends to be smaller there.

Figure 3 about here

In Figure 4 we have plotted the North-Eastern states of the US. The picture shows the (C)M(S)A's and all highways and railroads. The numbers in the map represent the CPS aggregated $\gamma_w' s$ (which will be used in the next section). Areas with lots of traffic connections like Boston, Chicago, Detroit and NYC have much smaller $\gamma_w' s$ than for example the rural parts of Tennessee, and Iowa.

4 Application: Estimation of a wage equation

In this section we look at the effect of our labor market density measure on wages. This application merely serves as an illustration. We do not have a structural interpretation of our estimation results per se. We put forward the simple hypothesis that wages are correlated with labor market density, for example by cost of living differentials, and we are just interested in what fraction of the variances in wages which is explained by regional factors can be attributed to labor market density. Hence, our results are a proof by implication: if density matters, it should pick up a substantial part of the cross-regional variation in wages. First, the following equation is estimated by OLS on 1991 CPS data:

$$\begin{aligned} \log w_{ij} &= \alpha_1 + \beta_1 X_1 + \lambda \gamma_j + \varepsilon_{1ij} \\ SSR &= 20562.56, R^2 = 0.3509 \end{aligned} \tag{5}$$

Where $\log w_{ij}$ is the log (gross) hourly wage of worker i from region j and X_1 contains all the standard variables of the wage equation[5]. The coefficient λ (with t-value) is : -0.39 (36.90). Compared to the female, -0.19 (17.23), and black -0.08 (10.72) dummies, this is a huge effect. Next we are interested in the extra variance of wages that can be explained by regional differences and which fraction of this is taken care of by our density measure. Consider therefore the following two regressions:

$$\begin{aligned} \log w_{ij} &= \alpha_2 + \beta_2 X_1 + \varepsilon_{2ij} \\ SSR &= 20985.44, R^2 = 0.3375 \end{aligned} \tag{6}$$

$$\begin{aligned} \log w_{ij} &= \beta_3 X_1 + \chi R_j + \varepsilon_{3ij} \\ SSR &= 19927.505, R^2 = 0.3662 \end{aligned} \tag{7}$$

[5]As explanatory variables we took: a constant, female, unmarried, female*unmarried, and black dummies, dummies for completed education (12, 14, 16, 18 years), education (yrs), cubic polinomial in experience and experience*education,female*experience, female* not married, female*not married* experience, $N = 66211$.

Where X_1 contains all the standard variables of the wage equation which we discussed before, R_j is a set of 49 state (we excluded Alaska and Hawaii) and 126 (C)M(S)A dummies (for each possible (C)M(S)A state combination there exists a unique γ_w).

We can conclude from those equations that regional effects account for 4.3% of the unexplained variance of wages and that our density measure explains 46.7% of this extra variation, which is substantial.

Finally, we tested how well our measure performs compared to the people-per-square-mile-measure (ppsm). For this test we restrict ourselves to the 126 (C)MSA's because only for those areas we have exactly matching information on ppsm. The $R^{2\prime}s$ of equations: (5), (6) and (7) are respectively: 0.353 ($\lambda = -0.29(19.32)$), 0.346 and 0.362.[6] The equivalence of (5) with ppsm/10000 instead of γ_w gives us an R^2 of 0.350 ($\lambda_{ppsm} = 0.85$ (16.05)). In other words, regional dummies explain 3.5% of unexplained wage variance, of this additional variance, 31.4% is captured by γ_w while only 17.1% is captured by people per square mile.

5 Discussion

We have shown that we can give a meaningful structural labor market interpretation of the Ellison and Glaeser (1997) index of concentration. One strong assumption we made is that the decision where to work and where to live are made sequentially rather than simultaneously, which is often not the case. The large and significant effect that our density measure has on wages is however encouraging. In future work we plan to use the measure to test for differences in match quality and match surplus in dense and non dense labor markets and in addition we want to test whether displaced workers find new jobs faster in dense labor markets.

6 Literature

Burda, M. and S. Profit (1996), "Matching across space Evidence on mobility in the Czech Republic", *Labour Economics*, 3, 255-278.

Chow, G.C. (1985), *Econometrics*, Mc Graw-Hill.

Coles, M. and E. Smith (1994), "Marketplaces and matching, *CEPR discussion paper*, 1048, London.

Coles, M. and E. Smith (1998), "Cross-Section estimation of the matching function: Evidence from England and Wales, *Economica*, 63, 589-598.

[6]Including the population size of the CMSA hardly explains extra wage variation, it leaves the R^2 at 0.352. It does changes the value of λ in equation (5) from -0.29 to -0.33 (19.2) .

Diamond P.A. (1991), Aggregate demand management in search equilibrium, *Journal of Political Economy*, 89, 798-812.

Ellison G. and E.L. Glasear (1997) Geographic concentration in U.S. manufacturing industries: A dartboard approach, *Journal of Political Economy*, 105, no.5, 889-927.

Jaeger D.A., S. Loeb, S. Turner and J. Bound (1998), Coding geographic areas across census years: creating consistent definitions of metropolitan areas, *NBER working paper* 6772, Cambridge MA.

McFadden, D., (1973), Conditional logit analysis of qualitative choice behavior, in P. Zarembka ed., *Frontiers in Econometrics*. New York: Academic Press.

Teulings C.N. and P.A. Gautier (2000), "Finding the right man for the job: Increasing returns to search?", *Mimeo*, Tinbergen Institute.

Wasmer E. and Y. Zenou (1999), Does space affect search? a theory of local unemployment, *CEPR Discussion Paper*, 2157.

A Appendix

A.1 Proof *Proposition 1*

First, define:

$$G_w = \sum_h (s_{wh} - x_h)^2$$

Write p_h for $\frac{\bar{\pi}_{wh}}{\sum_j \bar{\pi}_{wj}}$ and p_h for $p_h p_h$ and use the law of iterative expectations to write the definition of $E(G_w)$ as:

$$E(G_w) = \sum_h E_p E\left[(s_{wh} - x_h)^2 | p_h\right]$$

Next, note that since $x_h$is the mean of $\frac{\bar{\pi}_{wh}}{\sum_j \bar{\pi}_{wj}} = p_h$, and by the formula for the conditional variance: $var(s_{wh} - x_h | p_h) = var(s_{wh}|p_h) = E\left\{[(s_{wh} - x_h) - E\left((s_{wh} - x_h)|p_h\right)]^2 |p_h\right\}$ $= E\left[(s_{wh} - x_h)^2 | p_h\right] - E\left[s_{wh} - x_h | p_h\right)]^2 \Rightarrow E\left[(s_{wh} - x_h)^2 | p_h\right] = var(s_{wh}|p_h) +$ $E\left[(s_{wh} - x_h | p_h)\right]^2$.[7] Therefore,

$$E(G_w) = \sum_h E_{p_h}\ var(s_{wh}|p_h) + E_{p_h}\left[s_{wh} - x_h | p_h\right]^2$$

[7]Alternatively, $E\left[\sum_h (s_h - x_h)\right]^2 = \sum_h E\left([(s_h - p_h + p_h - x_h)]\right)^2 = \sum_h E_p \sum_h var(s_h) + \sum_h \left[s_h - x_h | p\right]^2$

Use $s_{wh} \equiv \frac{1}{W}\sum_k u_{kwh}$ (where u_{kwh} is a dummy which equals 1 if worker k who holds a job in area w, lives in h and zero otherwise and W is the size of area w) and expanding variance terms gives:

$$E(G_w) = \sum_h E_{p_h} \left[\left(\frac{1}{W}\right)^2 var(\sum_h u_{kh}|p_h) \right] + E\left[s_{wh} - x_h|p_h\right]^2$$

Use the fact that when X has a Bernoulli distribution, its variance is $p_h(1-p_h)$ and note that $E\left[s_{wh} - x_h|p_h\right] = (p_h - x_h)$. and $E\left[s_{wh} - x_h|p_h\right]^2 = E\left[(p_h - x_h)^2\right]$. Hence,

$$E(G_w) = \sum_h E_{p_h} \left\{ \frac{1}{W^2} \sum_h p_h(1-p_h) + (p_h - x_h)^2 \right\} \tag{8}$$

According to the specifications of (2) (3), $E(p_h) = x_h$ and $E[(p_h - x_h)^2] = var(p_h) = \gamma_w(x_h - x_h)$. Together this implies that:

$$\begin{aligned} E\left(p_h - (p_h - x_h)^2\right) &= E\left(p_h - (p_h^2 + x_h^2 - 2p_h x_h)\right) = x_h - \gamma_w(x_h - x_h^2) \Rightarrow \\ E\left((p_h - p_h^2\right) &= x_h - E(2p_h x_h - x_h^2) - \gamma_w(x_h - x_h^2) = \\ x_h - (2x_h^2 - x_h^2) - \gamma_w(x_h - x_h^2) &= (1-\gamma_w)(x_h - x_h^2) \end{aligned}$$

Substitute the relation above in (8) and adding subscript w again, gives:

$$E(G_w) = \sum_h E_p \left[\left(\frac{1}{W}\right)^2 (1-\gamma_w)(x_h - x_h^2) + \gamma_w(x_h - x_h^2) \right] \tag{9}$$

$$= (1 - \sum x_h^2) \left[\left(\frac{1}{W}\right)^2 (1-\gamma_w) + \gamma_w \right] \simeq (1 - \sum x_h^2)\gamma_w \tag{10}$$

A.2 Definitions

- **MA:** a large population nucleus,together with adjacent communities that have a high degree of economic and social integration with that nucleus. Each MA must contain either a place with a minimum population of 50,000 or a Census Bureau-defined urbanized area and a total MA population of at least 100,000 (75,000 in New England). A MA comprises one or more counties (cities and towns in New England) that have close economic and social relationships with the central county. An outlying county must have a specified level of commuting to the central counties and must meet certain standards regarding metropolitan character, such as population density, urban population, and population growth.

In the CPS, two related (not necessarily mutually exclusive) related concepts (1990 definitions) are used:

- **MSA** : relatively freestanding and not closely associated with other MA's, typically surrounded by non-metropolitan areas; the title of an MSA contains the name of its largest city and up to two additional city names.
- **CMSA**: consolidated metropolitan area. MA of more than 1 million people which may include one or more large urbanized counties that demonstrate very strong internal economic and social links within a CMSA. An example of a large CMSA is New York-New Jersey-Long Island.

B Pictures

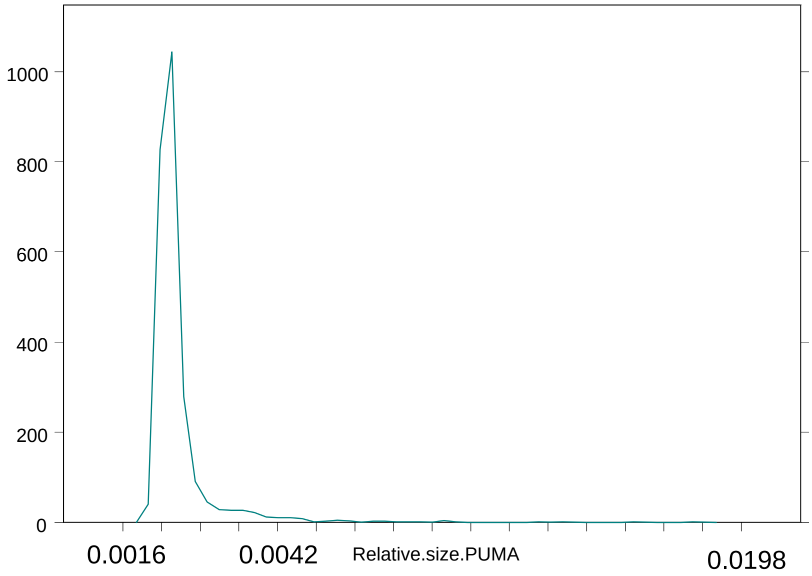

Figure 1: Density of area sizes, mean = 0.001, $\sigma^2 = 0.001$

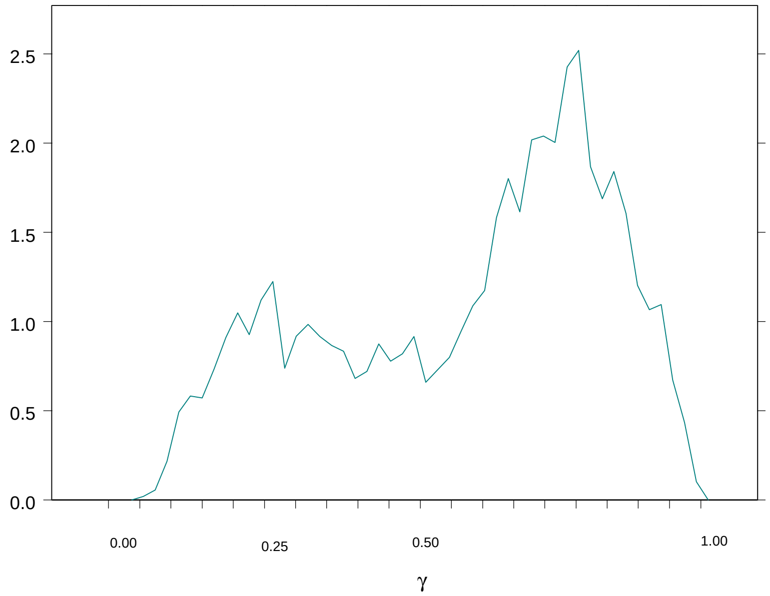

Figure 2: Density plot of γ from 1138 Census areas

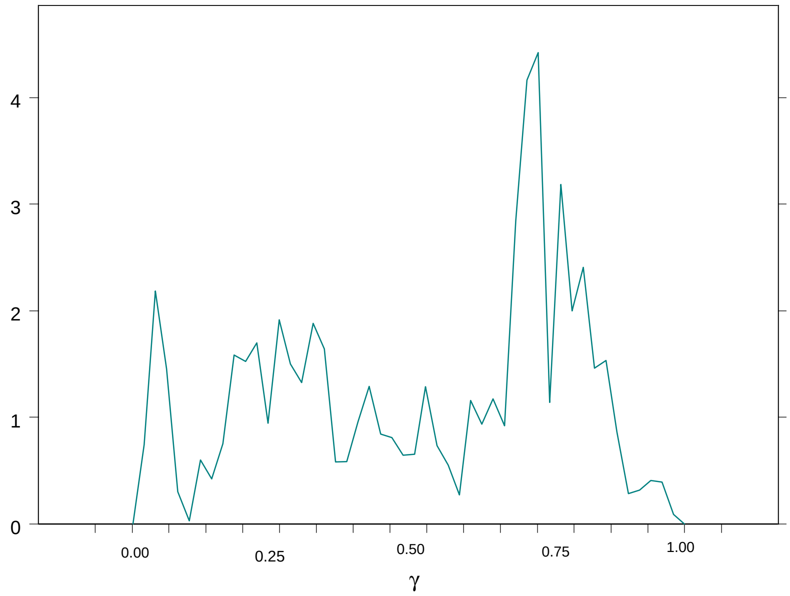

Figure 3: Density plot of γ from 182 CPS areas

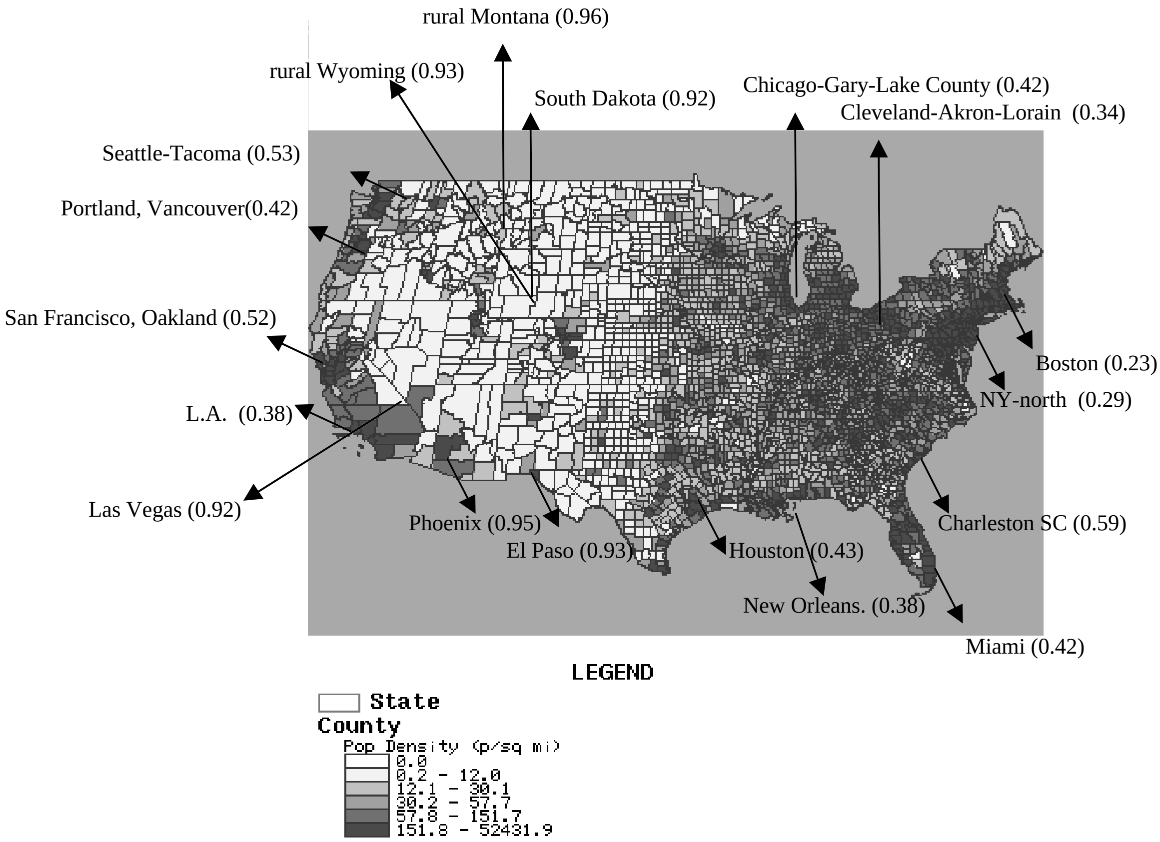

Figure 4: The relation between persons-per-square-mile and γ_{CPS}

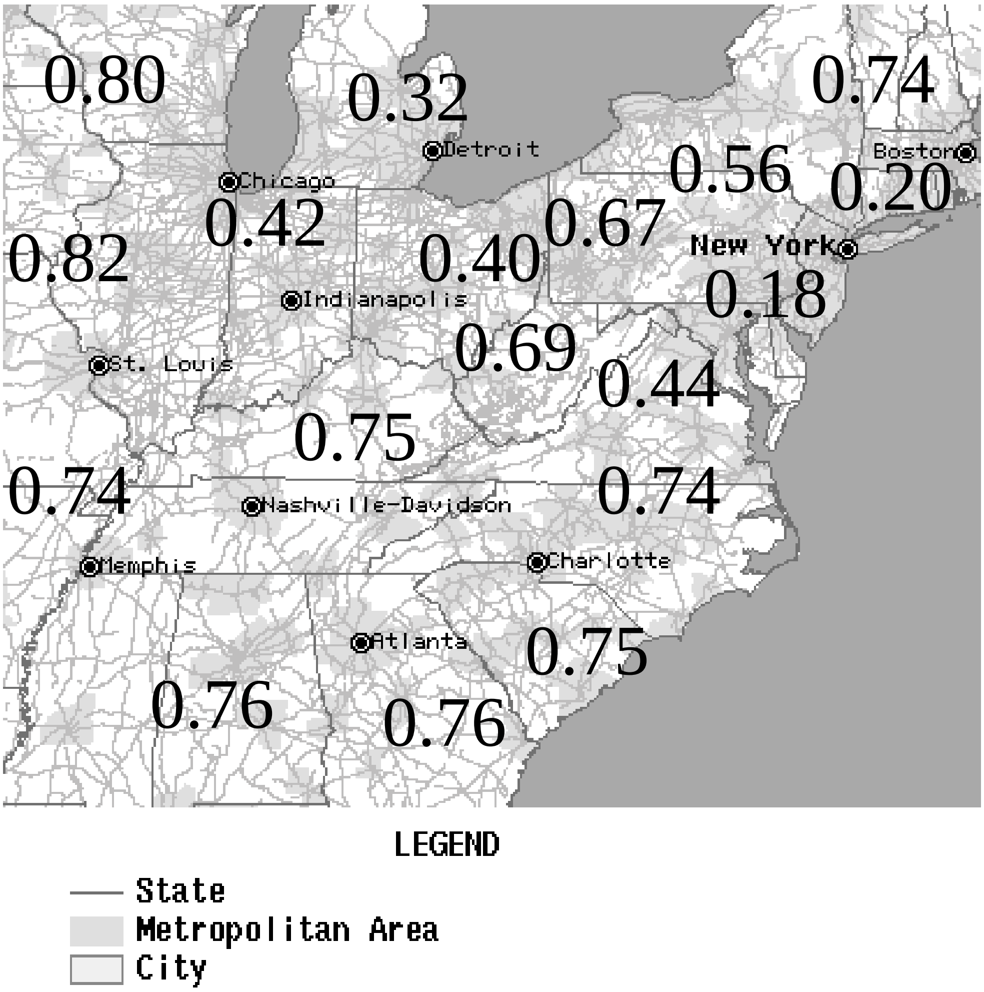

Figure 5: Highways, railroads and γ_{Census} for various (C)MSA's in the Nort Eastern states

C (C)MSA's/states ranked from dense to non dense

1 Washington DC Washington, 0.18201
2 Florida Orlando, FL 0.19529
3 Massachusetts Boston-Lawrence-Salem-Lowell-Brockton, MA 0.19993
4 Minnesota Minneapolis-St.Cloud MN (C) 0.21236
5 Connecticut Hartford-New Britain-Middletown-Bristol CT 0.26075
6 New Jersey Philadelphia-Wilmington-Trenton,NJ (C) 0.28888
7 Texas Dallas-Fort Worth, TX (C) 0.30358
8 Colorado Denver-Boulder, CO (C) 0.30739
9 Massachusetts Worcester, MA 0.31041
10 Connecticut New Haven-Meriden CT 0.31172
11 Michigan Detroit-Ann Arbor, MI (C) 0.31810
12 Rhode Island Providence-Pawtucket-Woonsocket, RI 0.31855
13 Georgia Atlanta, GA 0.32859
14 New York Buffalo-Niagara Falls, NY (C) 0.33560
15 Virginia Richmond-Petersburg, VA 0.34729
16 New York N.Y.-North. N.J.-Long Island, NY (C) 0.34776
17 Michigan Lansing-East Lansing MI 0.35209
18 Virginia Washington, VA 0.36757
19 Louisiana Baton Rouge, LA 0.37122
20 Tennessee Chattanooga, TN 0.37835
21 New York Albany-Schenectady-Troy, NY 0.37913
22 California Los Angeles city, CA 0.37934
23 Louisiana New Orleans LA 0.38304
24 Massachusetts Springfield, MA 0.38791
25 New York Syracuse, NY 0.38889
26 Kentucky Louisville, KY 0.39682
27 Maryland Baltimore, MD 0.40908
28 Michigan Grand Rapids MI 0.41137
29 Tennessee Knoxville, TN 0.41209
30 Florida Miami 0.41688
31 Oregon Portland OR (C) 0.41916
32 Illinois Chicago-Gary-Lake County, IL (C) 0.41966
33 Kentucky Cincinnati-Hamilton, KY (C) 0.42248
34 Missouri St. Louis, MO 0.42903
35 Maryland Washington, MD 0.43077
36 Texas Houston-Galveston-Brazoria, TX (C) 0.43306
37 Connecticut rural 0.43436
38 Virginia Norfolk-Virginia Beach-Newport News VA 0.43863

39 Michigan Flint, MI 0.44093
40 Illinois Rockford, IL 0.44260
41 North Carolina Fayetteville, NC 0.44308
42 Connecticut New London-Norwich, CT 0.44369
43 Pennsylvania Philadelphia-Wilmington-Trenton, PA (C) 0.44455
44 Kansas Kansas City KS 0.44475
45 North Carolina Greensboro-Winston-Salem-High Point, NC 0.45378
46 California Sacramento, CA 0.47082
47 California Modesto, CA 0.47128
48 Tennessee Memphis, TN 0.48094
49 Texas Beaumont-Port Arthur, TX 0.48404
50 Ohio Cincinnati-Hamilton, OH-KY-IN (C) 0.48589
51 Florida Melbourne-Titusville-Palm Bay, FL 0.49493
52 Washington Spokane, WA 0.49705
53 Pennsylvania Harrisburg-Lebanon-Carlisle, PA 0.49721
54 Missouri Kansas City MO-KS 0.49750
55 Indiana Fort Wayne, IN 0.49753
56 South Carolina Columbia, SC 0.50242
57 California Fresno, CA 0.50661
58 New York Rochester, NY 0.50816
59 Texas Austin, TX 0.51424
60 California San Francisco-Oakland-San Jose, CA (C) 0.51455
61 South Carolina Augusta, GA-SC 0.51538
62 Iowa Des Moines, IA 0.51545
63 California Bakersfield, CA 0.52003
64 Washington Seattle-Tacoma, WA (C) 0.53220
65 Mississippi Jackson, MS 0.53408
66 Indiana Indianapolis, IN 0.53501
67 Wisconsin Madison, WI 0.53776
68 Tennessee Nashville, TN 0.54252
69 Oregon Eugene-Springfield, OR 0.54305
70 Illinois Peoria, IL 0.54461
71 Pennsylvania Allentown-Bethlehem, PA-NJ 0.55004
72 Massachusetts rural 0.55970
73 Kentucky Lexington-Fayette, KY 0.56318
74 Illinois Davenport-Rock Island-Moline, IA-IL 0.56678
75 Oklahoma Oklahoma City, OK 0.57660
76 Georgia Macon-Warner Robins, GA 0.57871
77 Ohio Youngstown-Warren, OH 0.58012
78 Nevada Reno, NV 0.58201
79 Ohio Dayton-Springfield, OH 0.58718
80 Georgia Chattanooga, TN-GA 0.58889
81 Nebraska Omaha, NE-IA 0.59182

82 Wisconsin Milwaukee-Racine, WI (C) 0.59276
83 South Carolina Charleston, SC 0.59322
84 North Carolina Raleigh-Durham, NC 0.59538
85 Colorado Colorado Springs, CO 0.59608
86 Texas San Antonio, TX 0.59681
87 New Hampshire rural 0.59974
88 Wisconsin Appleton-Oshkosh-Neenah, WI 0.60155
89 North Carolina Charlotte-Gastonia-Rock Hill NC-SC 0.60192
90 California Salinas-Seaside-Monterey, CA 0.61512
91 Ohio Toledo, OH 0.61773
92 Indiana Louisville, KY-IN 0.61824
93 Iowa Davenport-Rock Island-Moline, IA-IL 0.62349
94 Alabama Birmingham, AL 0.62628
95 Alabama Montgomery, AL 0.62861
96 Tennessee Johnson City-Kingsport-Bristol, TN-VA 0.62942
97 Michigan Saginaw-Bay City-Midland, MI 0.64866
98 West Virginia Huntington-Ashland, WV-KY-OH 0.65689
99 Ohio Columbus, OH 0.66128
100 South Carolina Greenville-Spartanburg, SC 0.66141
101 Indiana rural 0.66600
102 Pennsylvania Pittsburgh-Beaver Valley PA (C) 0.67295
103 Oklahoma Tulsa, OK 0.67564
104 Florida Sarasota, FL 0.67790
105 South Carolina Charlotte-Gastonia-Rock Hill NC-SC 0.68105
106 Maryland rural 0.69311
107 Ohio rural 0.69361
108 New York Binghampton, NY 0.69981
109 Ohio Canton, OH 0.70070
110 Utah Salt City-Ogden, UT 0.71502
111 Delaware rural 0.71702
112 Vermont rural 0.71874
113 Florida Jacksonville, FL 0.71967
114 Indiana Evansville, IN-KY 0.72104
115 Pennsylvania York PA 0.72320
116 Pennsylvania Reading, PA 0.72631
117 Pennsylvania Scranton-Wilkes-Barre, PA 0.72731
118 Virginia rural 0.72983
119 Pennsylvania rural 0.73174
120 Alabama rural 0.73288
121 Georgia Augusta, GA-SC 0.73401
122 New York Utica-Rome, NY 0.73627
123 North Carolina rural 0.74169
124 Maine rural 0.74180

125 Arkansas Little Rock-North Little Rock, AR 0.74541
126 South Carolina rural 0.74941
127 Tennessee rural 0.75257
128 Michigan rural 0.75505
129 New York rural 0.75518
130 West Virginia rural 0.75523
131 Illinois rural 0.75701
132 Louisiana rural 0.75717
133 Florida West Palm Beach-Boca Raton-Delray FL 0.75848
134 Georgia rural 0.76103
135 Florida Tampa-St. Petersburg-Clearwater, FL 0.76119
136 Mississippi rural 0.76361
137 Texas Killeen-Temple, TX 0.76615
138 Texas Corpus Christi, TX 0.76942
139 West Virginia Charleston, WV 0.78167
140 California Stockton, CA 0.78215
141 Missouri rural 0.79070
142 Minnesota rural 0.79186
143 Kentucky rural 0.79450
144 Florida rural 0.79725
145 Wisconsin rural 0.79860
146 Arkansas rural 0.80362
147 Alabama Mobile, AL 0.80454
148 Kansas Wichita, KS 0.80993
149 New Mexico rural 0.82036
150 Texas rural 0.82189
151 North Dakota rural 0.82542
152 Iowa rural 0.82800
153 Washington rural 0.83144
154 Florida Lakeland-Winter Haven FL 0.84019
155 Pennsylvania Lancaster, PA 0.84155
156 California rural 0.84434
157 California Santa Barbara-Santa Maria-Lompoc, CA 0.84485
158 Arizona rural 0.85060
159 Colorado rural 0.85206
160 Louisiana Shreveport, LA 0.85560
161 Nebraska rural 0.86141
162 Idaho rural 0.86383
163 Kansas rural 0.86384
164 California Visalia-Tulare-Porterville, CA 0.86957
165 Oklahoma rural 0.87242
166 Pennsylvania Erie, PA 0.87639
167 Florida Daytona Beach, FL 0.87903

168 Oregon rural 0.87992
169 Utah rural 0.89050
170 Utah Provo-Orem, UT 0.89271
171 Texas Brownsville-Harlingen, TX 0.89633
172 Florida Fort Myers-Cape Coral FL 0.89751
173 Florida Pensacola, FL 0.90873
174 South Dakota rural 0.91771
175 Nevada Las Vegas NV 0.92344
176 Texas McAllen-Edinburg-Mission, TX 0.92793
177 Wyoming rural 0.93019
178 Texas El Paso, TX 0.93164
179 California San Diego CA 0.94695
180 Arizona Phoenix, AZ 0.94707
181 Montana rural 0.94790
182 Arizona Tucson, AZ 0.95404