Why does using individual and geocoded data matter in urban economics?

EFGS 2016 Conference

Pierre-Philippe Combes

University of Lyon

Sciences Po

Individual and geocoded data matter

- Claim of the presentation:

  Access to individual and geocoded data sets have much increased research possibilities.

  - Individual data: Information for each firm, worker, household, housing transaction,...
  - Geocoded data: Exact longitude and latitude, not only a spatial classification.

Individual and geocoded data matter

- Main purpose of urban economics: To evaluate the role and impact of cities.
- Cities generate
  - Urban costs: Higher housing and land prices, congestion in local public goods (transport,...), criminality, labour cost,...
  - Urban gains: Productivity, income, technology, urban amenities,...
- Study of the city size impact:
  - By how much urban gains increase with city size?
  - By how much urban costs increase with city size?

  $\Rightarrow$ Are cities too small or too large?
- Answers to all these questions can be inaccurate if a proper methodology, and, before anything, proper data, is not used.

  Access to individual and geocoded data much increases research possibilities.

Measuring spatial concentration

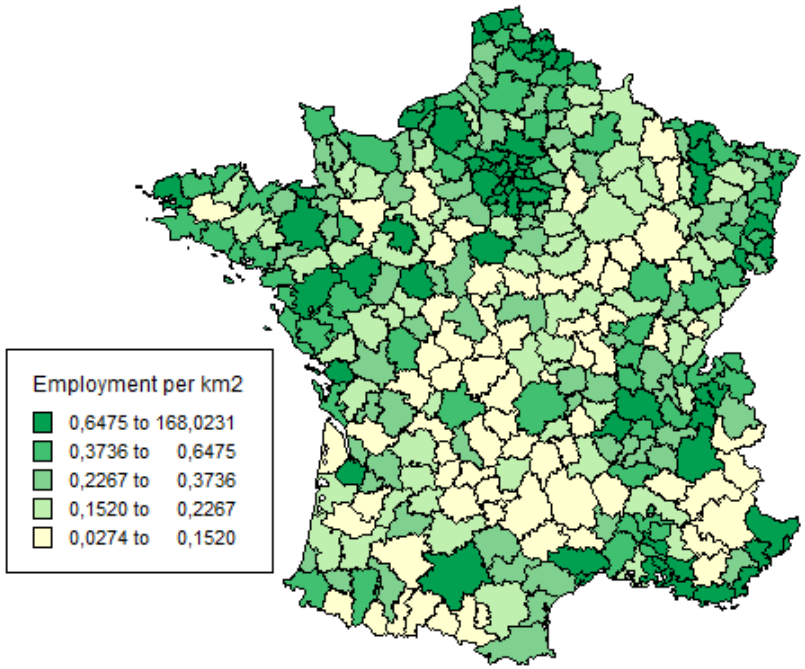
- Preliminary question: What is the degree of spatial concentration, ie do economic activities largely or only partly concentrate over space?
  - Are economic activities more concentrated spatially in the US or in Europe?
  - Is employment in electronics more or less concentrated in space than employment in chemical industries?
  - Is employment more or less concentrated than income?
- Possible to simply watch maps of corresponding economic variables but in general comparing concentration by simple visual inspection is difficult.
- Let us first recall some specificities of spatial data.

Spatial data

- Formerly, one needed to start with a spatial classification for all locations in a given area (a country or a continent for instance).
- For example, INSEE, the French National Institute for Statistics, designed a spatial classification named "employment areas".
  - The purpose was to capture the notion of "local labour markets".
  - Based on daily commuting patterns such that maximising the number of people living and working in the same employment area.
  - Named "Travel to Work Area" (TTWA) in some other countries.

French employment areas

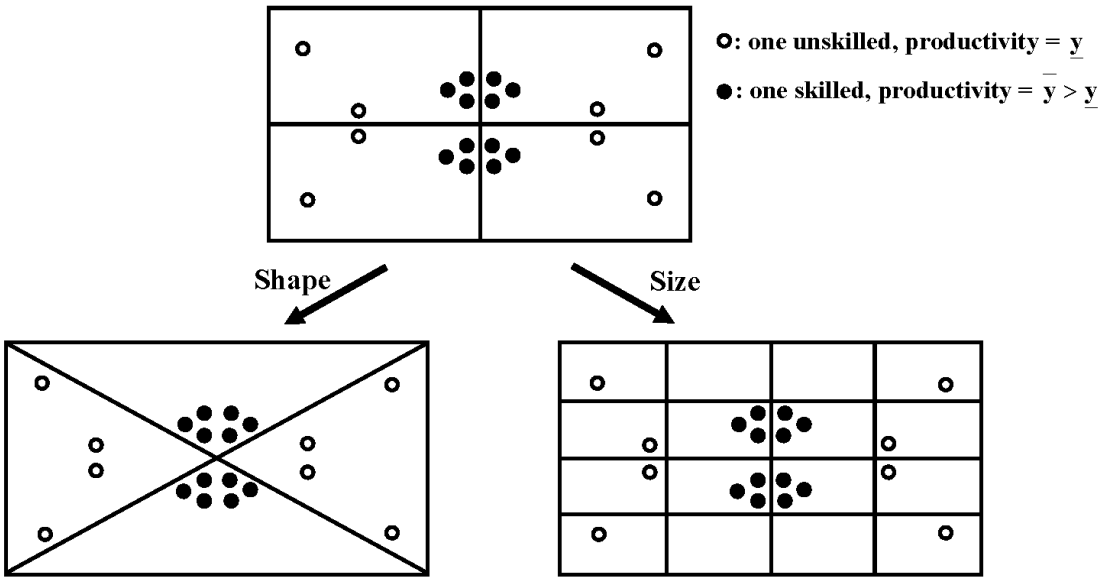
- This gives the following boundaries where ultimately around 80% of employees do work in the same employment area as the one where they live.



Employment density across French employment areas in 2007

The MAUP

- But choosing a specific spatial classification is not innocuous and may affect the measure of the economic phenomenon under study.
- This issue is called the Modifiable Areal Unit Problem (MAUP):



The size and shape sources of MAUP.

[Source: Briant, Combes, and Lafourcade, 2010]

The MAUP

- Econometric tools can reduce statistical biases due to the MAUP [Briant et al., 2010].

  This is because MAUP corresponds to a measurement error problem, to which econometrics methods can be robust.

  (as long as the model is correctly specified and errors are randomly distributed over locations)

- In general, better to choose a relevant spatial classification

  (TTWAs if one works on local labour markets for instance) ...

  ... to the extent that is does not generate the phenomenon one studies, see the MAUP example above...

Standard spatial concentration indexes

- Once a spatial classification is chosen, standard spatial concentration index are based on the level of activity (employment or production for instance) of industry s in area c, x_c^s, relatively to its national level, x^s: $\lambda_c^s = \frac{x_c^s}{x^s}$.
- Many examples adapted from inequality measures between individuals:
  - Isard index: $I^s = \frac{1}{2}\sum_{c=1}^{C} |\lambda_c^s - \lambda_c| \in [1/\lambda_{min}, 1]$.
  - Herfindhal index: $H^s = \frac{1}{C}\sum_{c=1}^{C} \lambda_c \left(\frac{\lambda_c^s}{\lambda_c}\right)^2 \in [1/C, 1]$.
  - α-Entropy index: $E^s(\alpha) = \frac{1}{\alpha^2-\alpha}\left[\sum_{c=1}^{C} \lambda_c \left(\frac{\lambda_c^s}{\lambda_c}\right)^\alpha - 1\right]$.
  - Gini index: $G^s = 1 - \sum_{n=1}^{C} \left[\lambda_{c(n)} - \lambda_{c(n-1)}\right]\left[\lambda_{c(n)}^s + \lambda_{c(n-1)}^s\right]$, once regions have been reordered according to $\frac{\lambda_c^s}{\lambda_c}$.
- Beyond the MAUP, they do not consider the distance between units. Invariant up to a permutation of units!

Continuous spatial concentration measures

- Geocoded data allows us to fully abstract from any spatial classification. Much more powerful.

  Proposed by Duranton and Overman [2005].

- Their approach is based on the exact location of plants and the distance between these plants.

The Duranton-Overman continuous approach: Data

(a) Basic Pharmaceuticals (SIC2441)

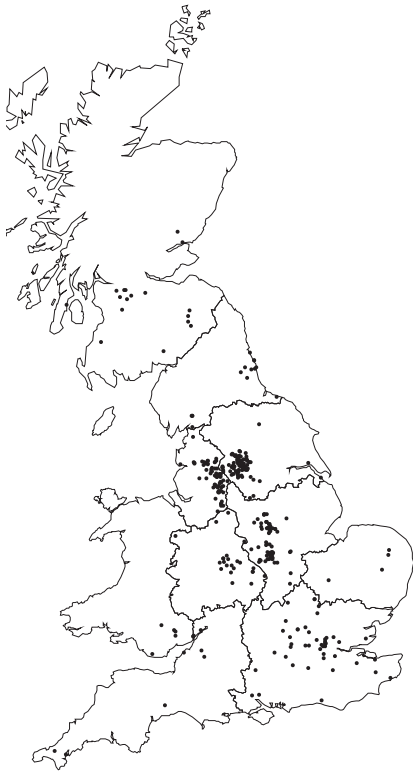
(b) Pharmaceutical Preparations (SIC2442)

Maps of the distribution of firms in two industries in the United Kingdom

[Source: Duranton and Overman, 2005]

The Duranton-Overman continuous approach: Data

(c) Other Agricultural and Forestry Machinery (SIC2932)



(d) Machinery for Textile, Apparel and Leather Production (SIC2954)

Maps of the distribution of firms in two industries in the United Kingdom

[Source: Duranton and Overman, 2005]

Examples of distance distribution

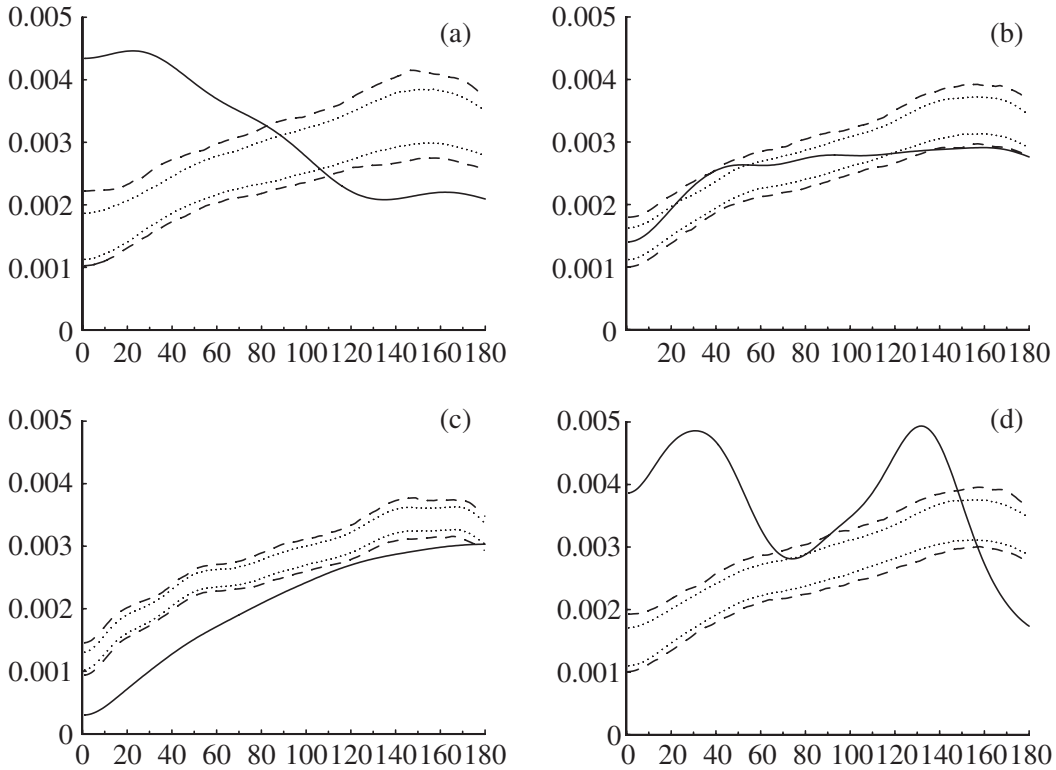


Figure 10.4. Densities and confidence intervals for four industries in the United Kingdom: (a) basic pharmaceuticals; (b) pharmaceutical preparations; (c) other agricultural and forestry; (d) machinery for textile, apparel, and leather production. (Source: Duranton and Overman (2005).)

Densities and confidence intervals for four industries in the United Kingdom

[Source: Duranton and Overman, 2005]

The Duranton-Overman continuous approach

- The difficulty consists in evaluating when the distance observed is larger than what would arise under random location choices of firms.
- Bootstrap approach:
  - The same number of plants as there are in the industry are randomly assigned to one of the possible locations (ie one occupied by at least one plant).
  - Then distances separating these plants are computed.
  - This is repeated 1000 times, which allows computing for each distance a mean value and a confidence interval around this mean.
  - When the observed distance is out this confidence interval, this mean that extra- or under- concentration is observed at this distance.

$\Rightarrow$ Distance between locations is taken into account (and is the key element of the measure), no MAUP anymore: Big improvement by using geocoded data.

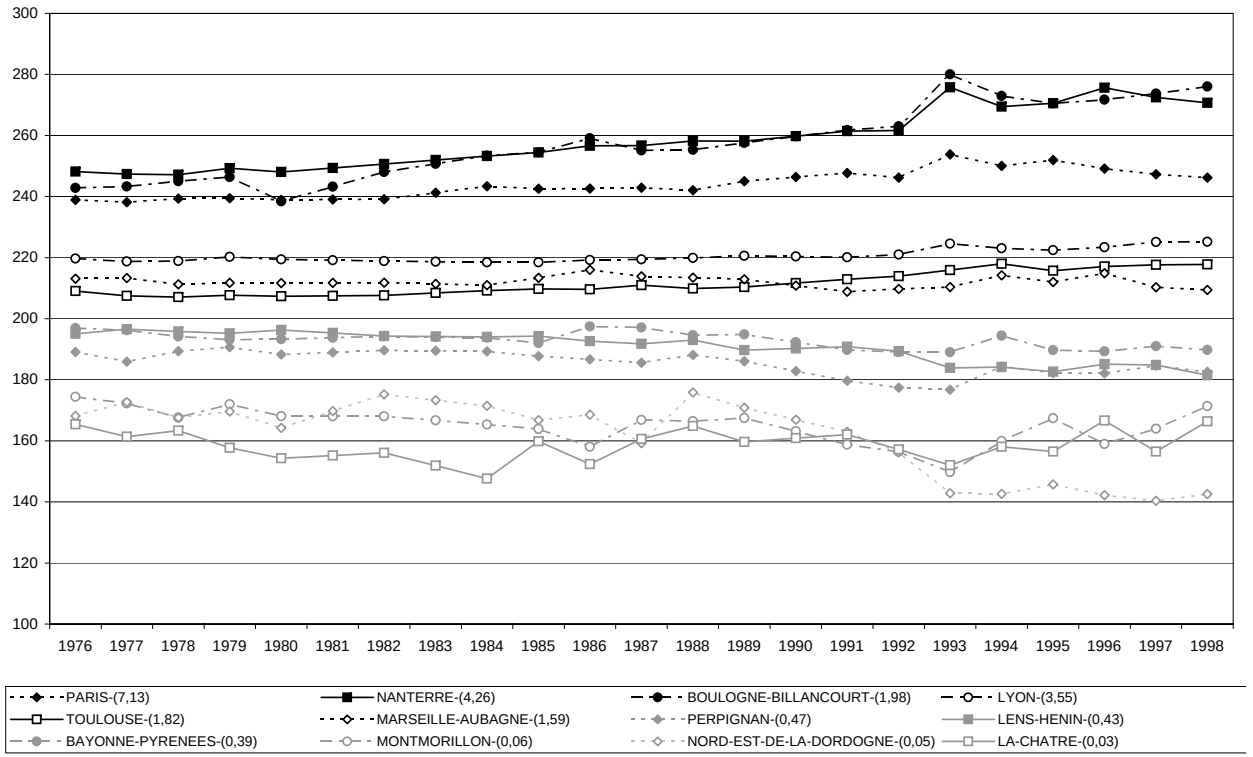
City size and productivity/income

- Urban economics emphasises the presence of gains from spatial concentration and cities:
  - Quantity and quality of matches between employers and employees,
  - Lower unit costs due to increasing returns to scale,
  - Knowledge creation and diffusion through local spillovers.

$\Rightarrow$ As a result, productivity, and then nominal income, increases with city size.

The French case

- Typical average local wage pattern for French employment areas:



Average nominal wages (detrended) in various French employment areas

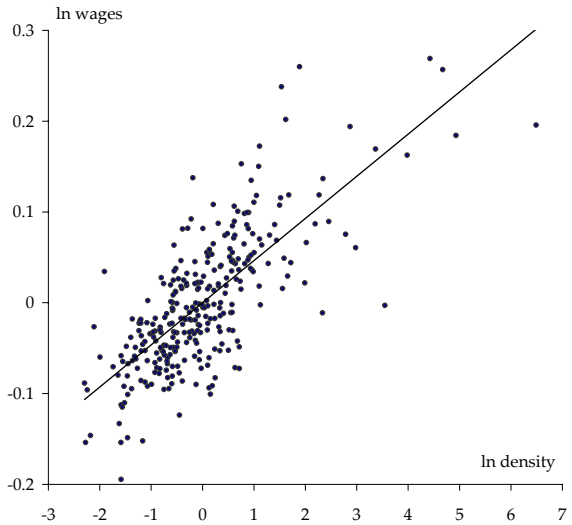
[Source: Combes, Mayer, and Thisse, 2008b]

The French case

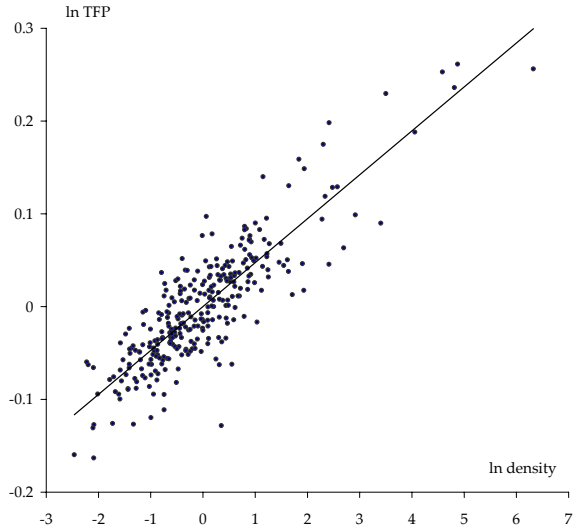
- Conclusions: Wages differences across local employment areas are:
  - Large,
  - Persistent,
  - Positively correlated with the area size.
- This is also found more systematically (ie on average for all employment areas in France) but also in many other contexts (over time, for other countries, etc).

The French case

- For instance, still for France, we have:



(a) Wages and employment density
(306 employment areas, 1976-1996 average)



(b) TFP (Olley-Pakes) and employment density
(306 employment areas, 1994-2002 average)

Wages and productivity against employment density

[Source: Combes, Duranton, Gobillon, and Roux, 2010]

⇒ Clear positive correlation between productivity and employment density (number of employees per square meter).

The productivity elasticity to density

- Big issue: The positive correlation can simply result from a pure composition effect, that more skilled workers locate more in larger cities.

$\Rightarrow$ Increasing city size would not increase income per capita a of given worker.

$\Rightarrow$ Impact of city size on income needs to be assessed for given skill level.

The productivity elasticity to density

- Literature was used to estimate on aggregate data:

$$\ln W_r = a + \beta \ln Density_r + \gamma HighSkilled_r + \varepsilon_r,$$

where

  - W_r: Regional average nominal wage,
  - $Density_r$: Regional employment density,
  - $HighSkilled_r$: Regional share of High Skilled workers,
  - ε_r: Random local productivity component (error term).

- β is the "elasticity" of wages (productivity) with respect to density.

  When density is 1% larger, productivity is β% larger.

The productivity elasticity to density

- Literature was used to estimate on aggregate data:

$$\ln W_r = a + \beta \ln Density_r + \gamma HighSkilled_r + \varepsilon_r,$$

where

  - W_r: Regional average nominal wage,
  - $Density_r$: Regional employment density,
  - $HighSkilled_r$: Regional share of High Skilled workers,
  - ε_r: Random local productivity component (error term).

- β is the "elasticity" of wages (productivity) with respect to density.

  When density is 1% larger, productivity is β% larger.

- Typical estimates for β obtained from aggregate data, in general based on spatial classifications considering not so small units,
  were between 0.06 and 0.10.

  But this was controlling for aggregate skills at the regional level only.

Using individual data and fine geography

- Combes, Duranton, and Gobillon [2008a] propose to use individual panel data to control for skills at the individual level.

  And to use a spatial classification corresponding to the scale where agglomeration mechanisms operate: Employment areas.

- French data set: Panel DADS (Déclarations Annuelles de Données Sociales)
  - Matched employer-employee data set
    $\Leftrightarrow$ Each observation: One person in one plant a given year.
  - Panel structure: Individuals followed over time (1976-2014) with location (municipalities that are aggregated in 341 employment areas) and industry (3-digit).
  - $1/24^{th}$ of salaried workers in the private sector: 18 millions of observations.
    $1/12^{th}$ and also public sector for recent years.
    Also cross-section version will all French salaried workers: 23 millions per year.
  - Earnings and costs, age, gender, skill level (occupation but not education).

Within individual estimation

- Such data allows studying the impact of density for a given individual who would move to another location.
- "Within-individual" estimation that controls for any individual skill constant over time:

  Worker's education but also parent's education, number of sisters and brothers, mobility of the family during childhood,...

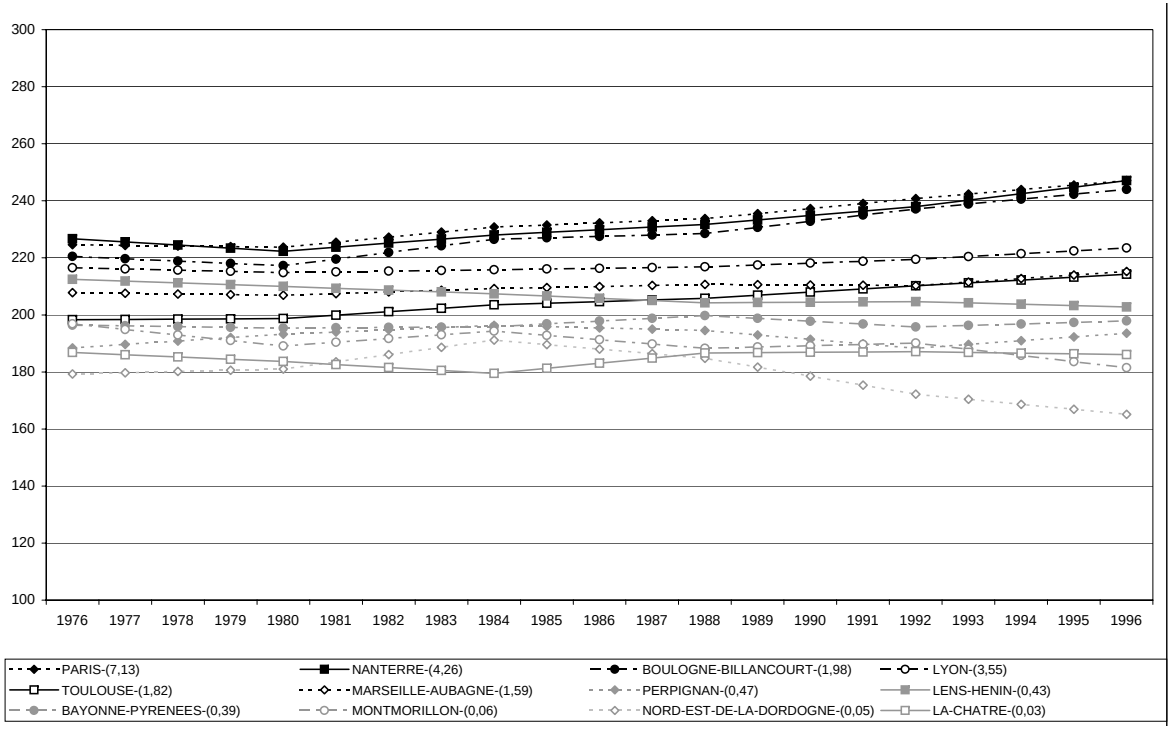
- Not a comparison of wage levels between workers but of the gap between the worker's current wage and the average wage over the career:

$$\ln w_{i,t} - \ln w_{i,\bullet} = \beta \left[\ln Density_{c(i,t),t} - \ln Density_{c(i,\bullet),\bullet} \right] + \nu_{c(i,t),t}.$$

  The average wage encompasses worker's skills, whatever their source.

$\Rightarrow$ The difference is net of any such effects, hence the impact of local city size on wages.

Wages differences net of skills effects



Average nominal wages net of individual effects (detrended)

$\Rightarrow$ Skills explain 40% to 50% of spatial wage disparities.

- A correlation with city size remains but it is lower.
  Typical estimates for β are now three times lower, at 0.02-0.03 (France, UK, Spain, Italy, Germany,...).

Implications of French estimates for a project as the Grand Paris

- Do such a difference (from 0.06 to 0.02) in estimates matter? YES!
- One of "Grand Paris" goals is to increase employment around Paris by much improving and extending the transport network.

  Typically, employment could increase by $500,000$ employees (an estimation not discussed here).

  What are the productivity gains expected from this increase?
- There are two sources of gains:
  - Density increases in the employment areas of the Grand Paris.

    All employees there experience an increase in productivity.
  - New employees come from less dense areas than the Grand Paris.

    They become more productive also.

Implications of French estimates for a project as the Grand Paris

- Let us assume $\beta = 0.02$.
- The density increase for employees already located in the GP case is small ($+9.6\%$), because density is already very high in the GP and adding $500,000$ employees is not that large.

  $\Rightarrow$ Productivity increases by only $1.096^{0.02} - 1 = 0.18\%$.
- But this applies to a very large number of workers (5.2 millions) and a very large yearly GDP (510 billions €).

  The resulting GDP gain is around 1 Billion € per year.

Implications of French estimates for a project as the Grand Paris

- The density variation for arriving employees is large: Density is multiplied by around 24 if they arrive from the average location in the rest of France.

  $\Rightarrow$ Their productivity increases by $24^{0.02} - 1 = 6.6\%$.

  Which now applies to "only" $500,000$ employees (who were also less productive than employees in the Grand Paris).

  Due to the large productivity gains, the GDP gain for these employees is around 2 billions € per year.

$\Rightarrow$ The overall gain is at around 3 billions a year,

given that overall project costs are at the moment at around 35 billions.

Implications of French estimates for a project as the Grand Paris

- But if one uses $\beta = 0.06$, figures are much larger.
- 2.8 billions € for the first gain, 6.6 billions € for the second gain.
- More than 9 billions € a year.

$\Rightarrow$ Makes a huge difference for policy makers and policy decision.

$\Rightarrow$ Controlling for individual effects and precise location is crucial.

The costs of agglomeration

- Part of nominal income gains for workers are absorbed by higher housing costs due to larger city size.
- Urban economics also attemps to evaluate the magnitude of such costs of agglomeration (housing costs, transport congestion, pollution, criminality,...)
- Again, using individual and geocoded data is very useful.

Housing costs

- Illustration on French data of British real estate agents famous sentence:

  "What matter for housing prices is location, location, and location".

- Combes, Duranton, and Gobillon [2016] study the determinants of land prices in France,

  using individual transactions (204,656 observations)

  with location known at the municipality level (more than 36,000 spatial units).

Determinants of housing costs

- Three types of characteristics considered simultaneously, corresponding to three geographical levels:
  - Parcels' characteristics: Land area, shape of the parcel, serviced or not, sold by a real estate agent or not,...
  - Municipality characteristics: Distance to the city (metropolitan area) centre, consumption amenities (swimming pool, theatres,...), productive amenities (stations, airports,...), geography (distance to coast, climate, ruggedness,...)
  - Metropolitan area characteristics: Same amenities and geographical features as at the municipality level + size (population and land area) + past growth.

Determinants of land prices

- Share of the Variance explained for different combinations of the variables.

Group of variables				
Parcel characacteristics	X	X		X
Municipality characteristics		X		X
City characteristics			X	X
Explained	47%	77%	52%	83%

  - Parcel characteristics alone explain 47% of the variance.
  - Parcel and municipality characteristics together explain 77% of the variance.
  - But city characteristics alone explain 52% of the variance.
  - 83% all together.

Determinants of land prices

- Share of the Variance explained for different combinations of the variables.

Group of variables				
Parcel characacteristics	X	X		X
Municipality characteristics		X		X
City characteristics			X	X
Explained	47%	77%	52%	83%

  - Parcel characteristics alone explain 47% of the variance.
  - Parcel and municipality characteristics together explain 77% of the variance.
  - But city characteristics alone explain 52% of the variance.
  - 83% all together.

$\Rightarrow$ Ignoring one or the other set of characteristics means ignoring a large share of land price determinants, which could bias the results.

- Note: Due to non-geocoded data, distances are still imperfectly measured.

  Hopefully, the geocoded version of the data set will be available soon to improve that!

Individual and geocoded data matter

- These are only three example of exercises in urban economics
  but in many other ones individual and geocoded data allow researchers to much refine their empirical strategy.
- Following individuals and locations over time is crucial too,
  in order to estimate impacts on time variations and not only cross-section.
- Merging and making consistent data of different nature,
  administrative, geospatial, mobile/"positioning apps" (Google, Waze,...),
  also opens many possibilities.

  Example of French BD Topo (IGN),
  about all French buildings with metric precision:
  - Combines satellite imagery and the land/housing administrative registry,
  - Improves the vertical precision (3D information),
  - Opens many possibilities in terms of building use and internal structure.

Anthony Briant, Pierre-Philippe Combes, and Miren Lafourcade. Does the size and shape of geographical units jeopardize economic geography estimations? *Journal of Urban Economics*, 67(3):287–302, 2010.

Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, 63(2):723–742, 2008a.

Pierre-Philippe Combes, Thierry Mayer, and Jacques-François Thisse. *Economic Geography: The integration of Regions and Nations*. Princeton University Press, New Jersey, 2008b.

Pierre-Philippe Combes, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. Estimating agglomeration effects with history, geology, and worker fixed-effects. In Edward L. Glaeser, editor, *Agglomeration Economics*, pages 15–65. Chicago University Press, Chicago, IL, 2010.

Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. The costs of agglomeration: House and land prices in French cities. Discussion Paper 9240, Centre for Economic Policy Research, 2016.

Gilles Duranton and Henry G. Overman. Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4):1077–1106, 2005.