Production Clusters

Thomas Chaney

Toulouse School of Economics

May, 2014

PRELIMINARY AND INCOMPLETE

Motivation

- Firms are *legal* entities.
- Production function has to do with *technology*.
- Firm boundaries often do not coincide with technological boundaries.
- **Questions:**
  1. How to define technological units?
  2. What are the properties of technological units?

Motivation

- Novel evidence: few shipments within firms.
    - Atalay et al. (AER 2014): US plants.
    - Ramondo et al. (2014): US multimationals
- If production chains not contained within firms, then where?
- **What this paper aims at:**
  Empirical analysis of micro production chains.

Input-output linkages

- I use input-output linkages to define clusters.
- Cohen & Frazzini (JF 2008): customers matter for firms profits.
- Kelly et al. (2013): customers affect firm's volatility.
- Barrot & Sauvagnat (2014): suppliers affect firms growth.
- Acemoglu et al (ECMA 2012): input-output matter at the macro level.

What I do

1. Define "production clusters" from the bottom-up.
2. Analyze the properties of those clusters.
3. Contrast firms and production clusters.

Main findings

1. Cluster sizes follows Pareto distribution (better than firms).
2. Upper tail of cluster size distribution tightly follows Zipf's law.
3. Clusters volatility decreases fast with size (unlike firms).

Data

Data

- Sample: publicly traded US firms (COMPUSTAT).
- Shipments of intermediates:
  - any customer accounting for more than 10% of sales is reported.
  - identity of supplier and customer (Cohen & Frazzini, JF 2008).
  - Atalay et al. (PNAS 2011): little bias from 10% cut-off rule.
- Total firm sales; total employment (unreliable).

"Distance" between 2 firms

- For all firms (i, j) in COMPUSTAT,

$$S_{i,j} = \text{dollar value of shipments from } i \text{to } j$$

- For all (i, j),

$$S_{i,\cdot} = \sum_j S_{i,j} = \text{shipments sent by } i$$
$$S_{\cdot,j} = \sum_i S_{i,j} = \text{shipments received by } j$$

- Measure of input-output "distance" between 2 firms,

$$d_{i,j} = 1 - \frac{S_{i,j} + S_{j,i}}{S_{i,\cdot} + S_{\cdot,j}} \in [0, 1]$$

"Distance" between 2 firms

- For all firms (i, j) in COMPUSTAT,

$$S_{i,j} = \text{dollar value of shipments from } i \text{to } j$$

- For all (i, j),

$$S_{i,\cdot} = \sum_j S_{i,j} = \text{shipments sent by } i$$
$$S_{\cdot,j} = \sum_i S_{i,j} = \text{shipments received by } j$$

- Measure of input-output "distance" between 2 firms,

$$d_{i,j} = 1 - \frac{S_{i,j} + S_{j,i}}{S_{i,\cdot} + S_{\cdot,j}} \in [0, 1]$$

- Alternative "distance" measures,

$$1 - \frac{S_{i,j} + S_{j,i}}{S_{i,\cdot} + S_{j,\cdot}}, 1 - \frac{S_{i,j} + S_{j,i}}{S_{\cdot,i} + S_{\cdot,j}}, 1 - \max\left\{\frac{S_{i,j}}{S_{i,\cdot}}, \frac{S_{j,i}}{S_{j,\cdot}}\right\}, 1 - \max\left\{\frac{S_{i,j}}{S_{\cdot,j}}, \frac{S_{j,i}}{S_{\cdot,i}}\right\}$$

Data limitations

1. 10% cut-off rule.
   Atalay et al. (PNAS 2011) suggest no systematic bias from it.

Data limitations

1. 10% cut-off rule.
   Atalay et al. (PNAS 2011) suggest no systematic bias from it.
2. Shipments sent (received from) outside COMPUSTAT.
   Nothing I can do for now.
   Plant level data would solve that (future research).

Data limitations

1. 10% cut-off rule.
   Atalay et al. (PNAS 2011) suggest no systematic bias from it.
2. Shipments sent (received from) outside COMPUSTAT.
   Nothing I can do for now.
   Plant level data would solve that (future research).
3. Shipments sent within firms.
   Atalay et al. (AER 2014) suggest this is serious.
   Plant level data would solve that (future research).

2 clustering algorithms

1. Partition around medoids (PAM).
2. Firm clustering algorithm (FCA).

Partition around medoids (PAM)

Objective

Minimize the average distance of firms to their closest "medoid".

1. Choose K "medoids" (cluster centers): $k = 1, \cdots, K$.
2. Allocate firms to their closest "medoid".

PAM: pros and cons

- Pros:
  - Off-the-shelf algorithm (used for data mining).
  - Flexible choice for # of clusters (K).
  - All firms in a cluster are "close" neighbors.
- Cons:
  - favors "horizontal" clusters.
  - misses multiple stages of production.

Firm clustering algorithm (FCA)

Objective

Any pair of "neighbors" belong to the same cluster.

1. Choose a distance threshold $\bar{d}$.
2. Any 2 firms a distance $d < \bar{d}$ are "neighbors".
3. Clusters are maximally connected sets of firms (neighbor).

Firm clustering algorithm (FCA)

Objective

Any pair of "neighbors" belong to the same cluster.

1. Choose a distance threshold $\bar{d}$.
2. Any 2 firms a distance $d < \bar{d}$ are "neighbors".
3. Clusters are maximally connected sets of firms (neighbor).

Note: similar to Rozenfeld et al. (AER 2010), City Clustering Algorithm (CCA).

FCA: pros and cons

- Pros:
  - Flexible choice of distance threshold $\bar{d}$.
  - Allows for complex vertical production chains.
  - Direct measure of the complexity of production within cluster.
- Cons:
  - leaves many single firm clusters (if $\bar{d}$ small).
  - bundles together firms connected through long chains.

Cluster properties

No "natural" clusters (PAM)

- Criterion for natural clusters (optimal choice of K): difference between distance within and between clusters
- No optimal choice of clusters ("fractal").
- In general, similar properties for various K's and $\bar{d}$'s.

Cluster size

- Cluster size = external sales:

$$Size_k\left(t\right) = \sum_{i \in k} Sales_i\left(t\right) - \sum_{i,j \in k} Shipments_{i,j}\left(t\right)$$

Note: only partial information on internal sales.

Cluster size

- Cluster size = external sales:

$$Size_k(t) = \sum_{i \in k} Sales_i(t) - \sum_{i,j \in k} Shipments_{i,j}(t)$$

Note: only partial information on internal sales.

- Cluster size = total employment:

$$Size_k(t) = \sum_{i \in k} Employment_i(t)$$

Note: employment data in COMPUSTAT unreliable.

3 properties

1. Size distribution (Pareto, lognormal?).
2. Growth versus size (Gibrat's law?).
3. Volatility versus size (diversification, LLN?).

Size distributions

2-sided Pareto distribution for cluster sizes

- Cluster sizes are 2-sided Pareto distributed (p.d.f. f, c.d.f. F),

$$f(S) = \begin{cases} \frac{\alpha}{\beta(\alpha+\beta)S_0^{\beta}} S^{\beta-1} & \text{if } S \leq S_0 \\ \frac{\beta}{\alpha(\alpha+\beta)S_0^{-\alpha}} S^{-\alpha-1} & \text{if } S > S_0 \end{cases}$$

$$F(S) = \frac{\alpha}{\alpha+\beta}\left(\frac{S}{S_0}\right)^{\beta} \text{ if } S \leq S_0$$

$$1 - F(S) = \frac{\beta}{\alpha+\beta}\left(\frac{S}{S_0}\right)^{-\alpha} \text{ if } S > S_0$$

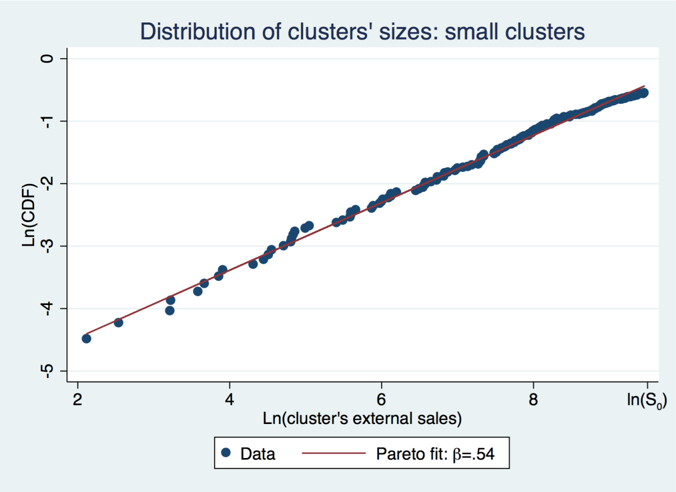
2-sided Pareto distribution for cluster sizes

- Cluster sizes are 2-sided Pareto distributed (p.d.f. f, c.d.f. F),

$$\ln(f(S)) = \begin{cases} constant + (\beta - 1)\ln S & \text{if } \ln S \leq \ln S_0 \\ constant - (\alpha + 1)\ln S & \text{if } \ln S > \ln S_0 \end{cases}$$

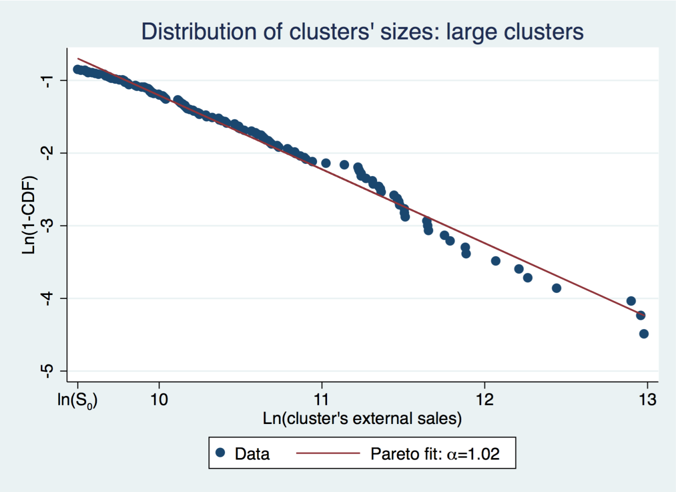
$$\begin{aligned} \ln(F(S)) &= constant + \beta \ln S \text{ if } \ln S \leq \ln S_0 \\ \ln(1 - F(S)) &= constant - \alpha \ln S \text{ if } \ln S > \ln S_0 \end{aligned}$$

2-sided Pareto for cluster sizes (FCA)



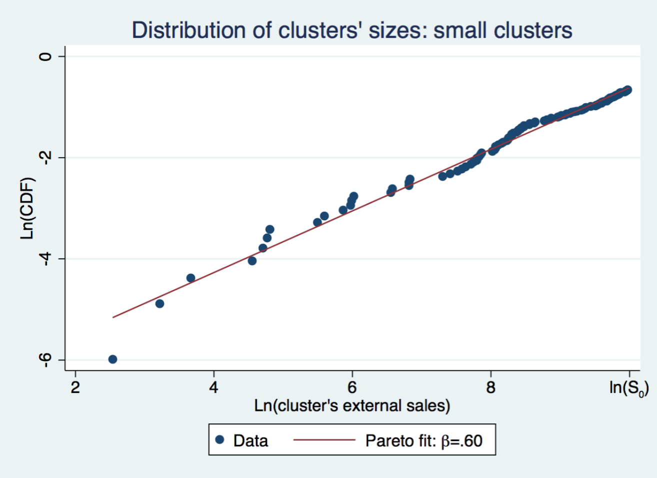
Source: COMPUSTAT 1986-1991. Firm Clustering Algorithm, $\bar{d} = .6$.

2-sided Pareto for cluster sizes (FCA)



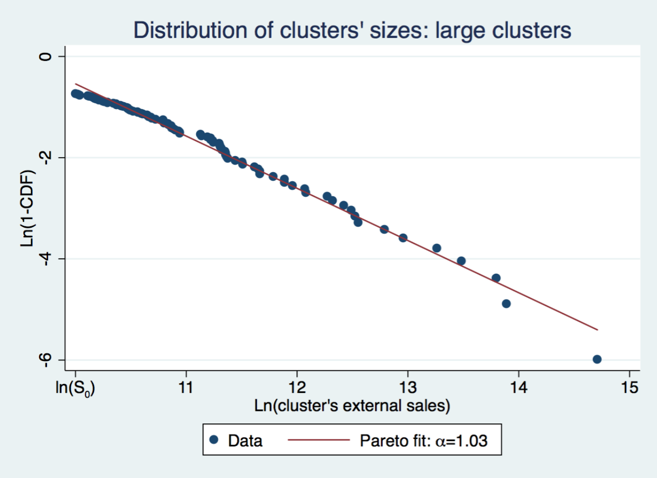
Source: COMPUSTAT 1986-1991. Firm Clustering Algorithm, $\bar{d} = .6$.

2-sided Pareto for cluster sizes (PAM)



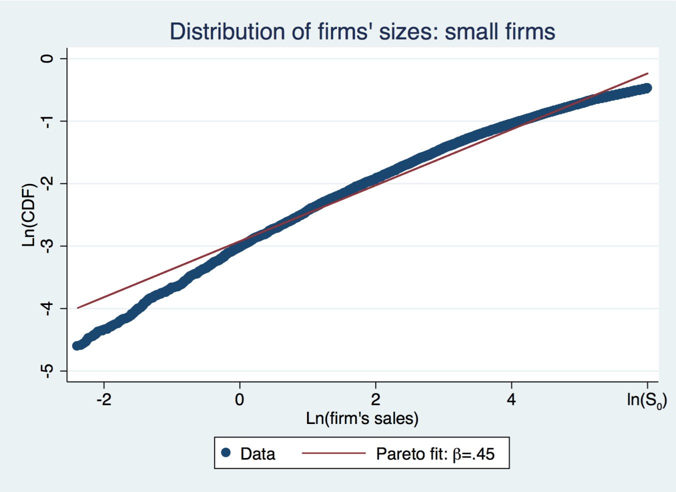
Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 200$.

2-sided Pareto for cluster sizes (PAM)



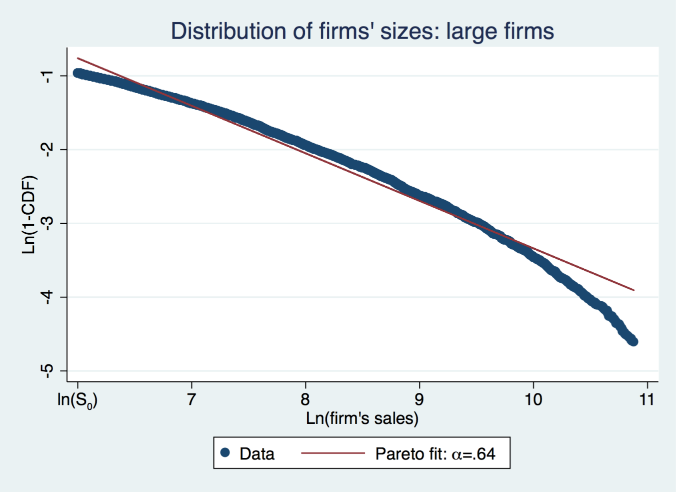
Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 200$.

so-so 2-sided Pareto for firm sizes



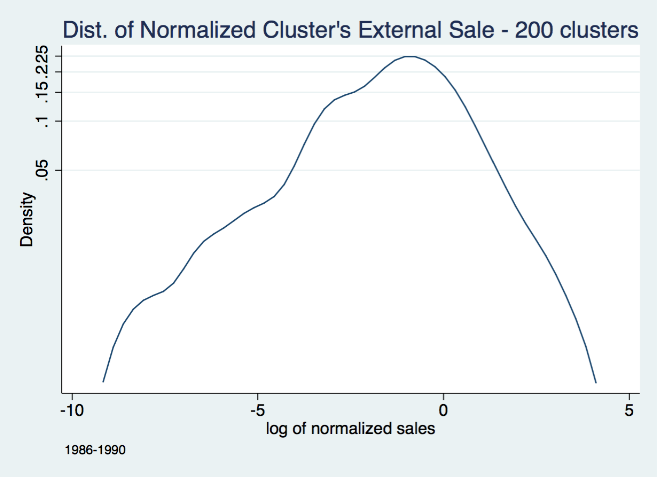
Source: COMPUSTAT 1986-1991.

so-so 2-sided Pareto for firm sizes



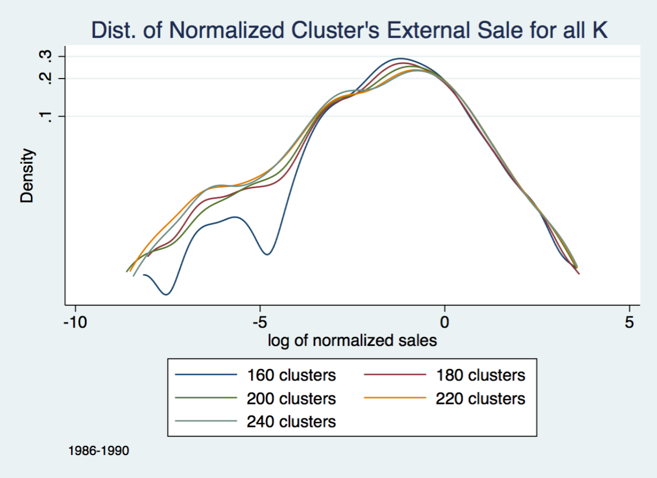
Source: COMPUSTAT 1986-1991.

2-sided Pareto for cluster sizes (PAM)



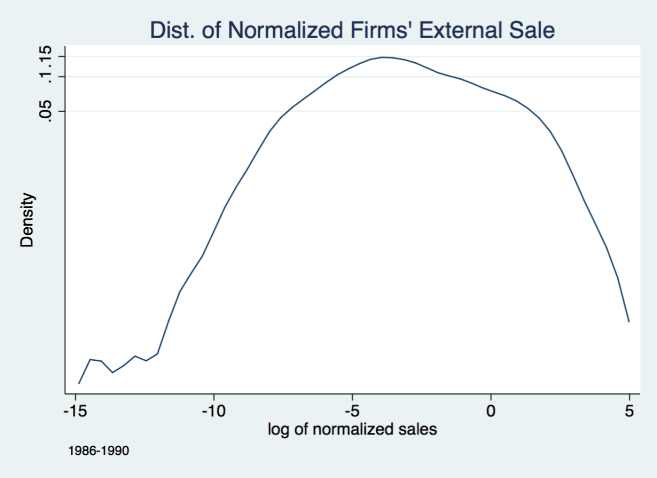
Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 200$.

2-sided Pareto for cluster sizes (PAM)



Source: COMPUSTAT 1986-1991. Partitioning Around Medoids.

so-so 2-sided Pareto for firm sizes



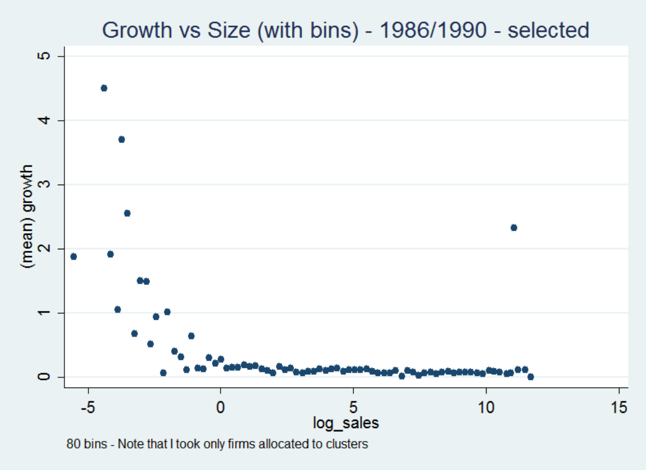
Source: COMPUSTAT 1986-1991.

Growth vs. size

Gibrat's law

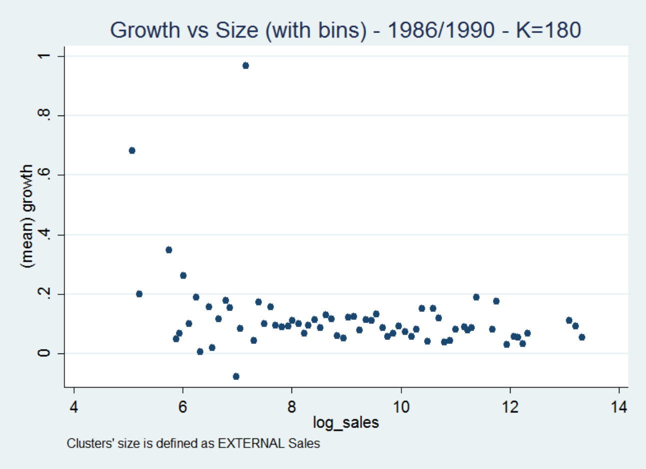
- For firms, growth rate $\approx$ independent of size.
- Departure from Gibrat's law for small firms (grow faster).
- Same for clusters.

Growth versus size: firms



Source: COMPUSTAT 1986-1991.

Growth versus size: clusters



Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 180$.

Volatility vs. size

Diversification

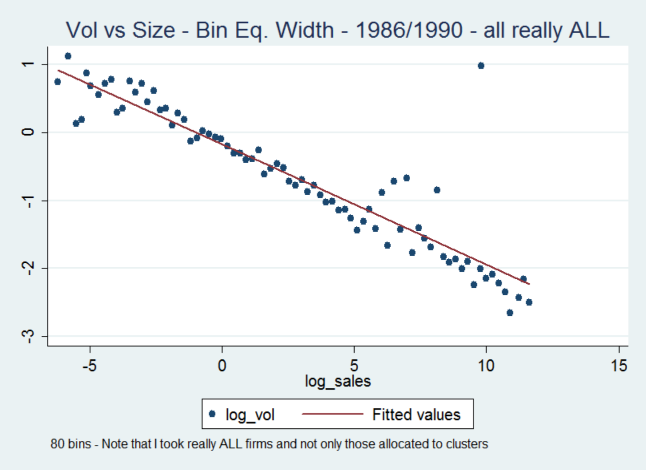
- For firms, volatility falls slowly with size,

$$s.d.\left(\gamma\left(S\right)\right) = constant\, S^{-\zeta}$$
$$with\, \zeta \approx .16$$

- For clusters, volatility falls faster with size,

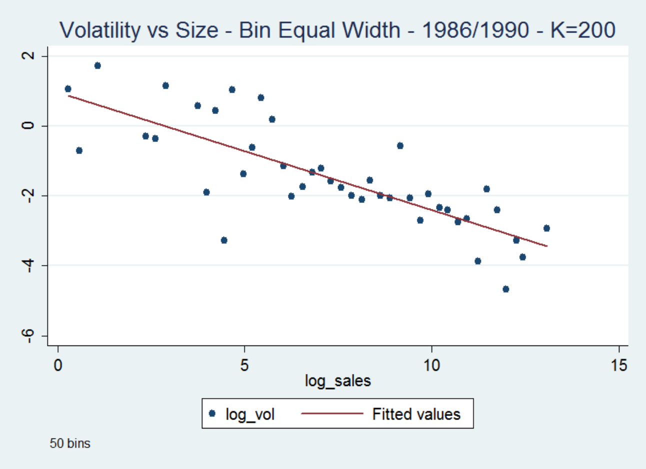
$$s.d.\left(\gamma\left(S\right)\right) = constant\, S^{-\zeta}$$
$$with\, \zeta \approx .33$$

Volatility versus size: firms



Source: COMPUSTAT 1986-1991.

Volatility versus size: clusters

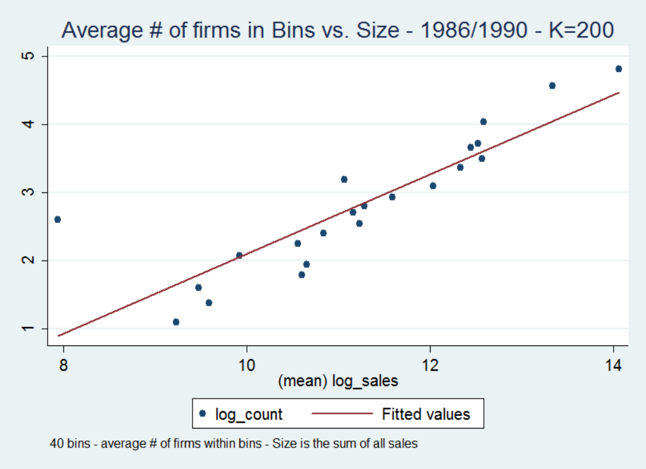


Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 200$.

Volatility versus size

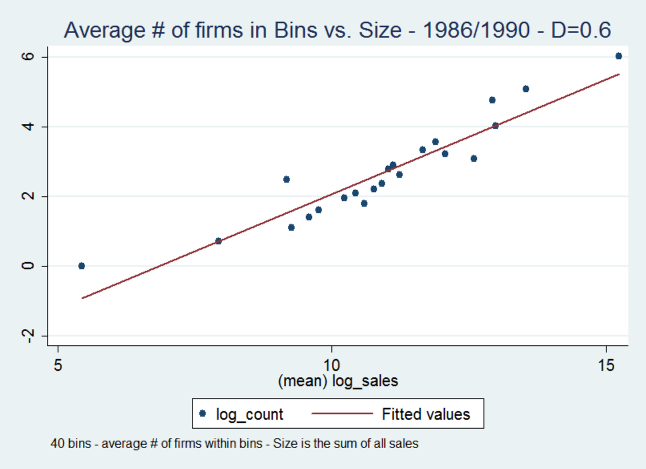
Dep. Var.: s.d.(growth rate)	All firms (1)	200 clusters (2)
ln (Sales)	-0.163	-0.337
	(.0016)	(.041)
intercept	-0.394	0.974
	(.098)	(.333)
N. obs	75	42
R^2	0.58	0.625

Larger clusters have larger and more firms



Source: COMPUSTAT 1986-1991. Partitioning Around Medoids, $K = 200$.

Larger clusters have larger and more firms



Source: COMPUSTAT 1986-1991. Firm Clustering Algorithm, $\bar{d} = .6$.

A simple explanation

1. **If** large clusters have both **larger and more firms**,

$$\bar{S}_{firms}\left(S_{cluster}\right) \propto \left(S_{cluster}\right)^{\alpha}$$

$$\#_{firms}\left(S_{cluster}\right) \propto \left(S_{cluster}\right)^{1-\alpha}$$

2. **If** larger firms are less volatile than smaller ones,

$$s.d.\left(\gamma_{firm}\left(S\right)\right) \propto S^{-\zeta_{firm}}$$

$$with\,\zeta_{firm} \approx \frac{1}{6}$$

3. **And if** firms within clusters are **independent**,
4. **Then**,

$$s.d.\left(\gamma_{cluster}\left(S\right)\right) \;\propto\; S^{-\zeta_{cluster}}\,with\,\zeta_{cluster} \;=\; \frac{\alpha\left(2\zeta_{firm}-1\right)+1}{2}$$

$$\left.\begin{array}{r} \alpha \approx \frac{1}{2} \\ \zeta_{firm} \approx \frac{1}{6} \end{array}\right\} \Rightarrow \zeta_{cluster} \approx \frac{1}{3}$$

Conclusion

Next steps

1. Plant level data:
    1. Turkey (VAT forms)
    2. Belgium
    3. other countries?
2. Sharing of intangibles within clusters:
    1. workers flows
    2. idea flows (?)
3. Theory:
    1. span of control of managers within firms
    2. search and matching for suppliers/customers

Thank you