RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Panel Data Methods and Applications

Online Course

DAY2 - November 20th, 2020

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n° 824091

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Introduction to VICO

VICO

RISIS

- **VICO** contains geographical, industry and accounting information on companies:
  - Founded starting from 1/1/1988
  - Which have received at least 1 VC investment starting from 1/1/1998
  - Operating in seven EU27 countries + United Kingdom + Israel

- **Strengths:**
  - Extent of information gathered
  - Overall number of companies: 24,238
  - Country coverage
  - Information on 8,568 distinct investors

Units and definition of observations

RISIS

- Data can be broadly classified as:
  - Company-level data
    - General company information
    - Accounting data
  - Investor-level data
  - Investment-level data (deal specific)

Company-level data
General company information

RISIS

- Company ID (Company ID code, Company Name)
- Address (e.g. Nation, City, Zip Code, NUTs, FUAs, Lat, Long)
- Industry classification (e.g. NACE Rev. 2 codes, NACE main section)
- Year of incorporation
- Status (active, listed, acquired, bankrupt)

Company-level data
Accounting data

RISIS

- Income statement (e.g Turnover, EBITDA, Net profit)
- Balance sheet (e.g. Total assets, Shareholder funds, Debt)
- Number of employees

Investor-level data

RISIS

- Investor ID (Investor ID code, Investor Name)
- Address (e.g. Nation, City, Zip Code, NUTs, FUAs, Lat, Long)
- Type of investor (Independent VC, Corporate VC, Bank-affiliated VC, Governmental VC; BA; Crowdfunding; Other)
- Year of incorporation of VC management company
- Industry classification (e.g. NACE Rev. 2 codes, NACE main section)

Investment-level data
Deal-specific information

RISIS

- Date of the investment
- Round number
- Identity of investors involved in the deal
- Total amount invested in the round

var62[1]

	CompanyID	RoundNumber	InvestorID	InvestorType	d_Undiscol~r
1	VICO1	1	VICOInvestor02026	IVC	.
2	VICO10	1	VICOInvestor01211	IVC	.
3	VICO10	2	VICOInvestor01211	IVC	.
4	VICO10	3			1
5	VICO10	4	VICOInvestor01394	BVC	.
6	VICO10	4	VICOInvestor01211	IVC	.
7	VICO100	1	VICOInvestor06246	IVC	.
8	VICO1000	1			1
9	VICO1000	1	VICOInvestor08215	IVC	.
10	VICO10000	1	VICOInvestor08131	IVC	.
11	VICO10001	1	VICOInvestor01565	IVC	.
12	VICO10002	1	VICOInvestor01279	IVC	.
13	VICO10002	2	VICOInvestor01279	IVC	.
14	VICO10002	3			1
15	VICO10002	4	VICOInvestor01279	IVC	.
16	VICO10003	1	VICOInvestor08448	BVC	.
17	VICO10004	1	VICOInvestor03328	IVC	.
18	VICO10004	2	VICOInvestor07545	IVC	.
19	VICO10004	2			1
20	VICO10004	3	VICOInvestor07545	IVC	.
21	VICO10004	3	VICOInvestor03657	IVC	.
22	VICO10004	3	VICOInvestor03328	IVC	.
23	VICO10004	3	VICOInvestor03640	BVC	.
24	VICO10005	1	VICOInvestor00731	IVC	.
25	VICO10007	1	VICOInvestor00699	IVC	.

Vars: 5 of 55 Order: Modified | Obs: 52.657

Data sources

RISIS

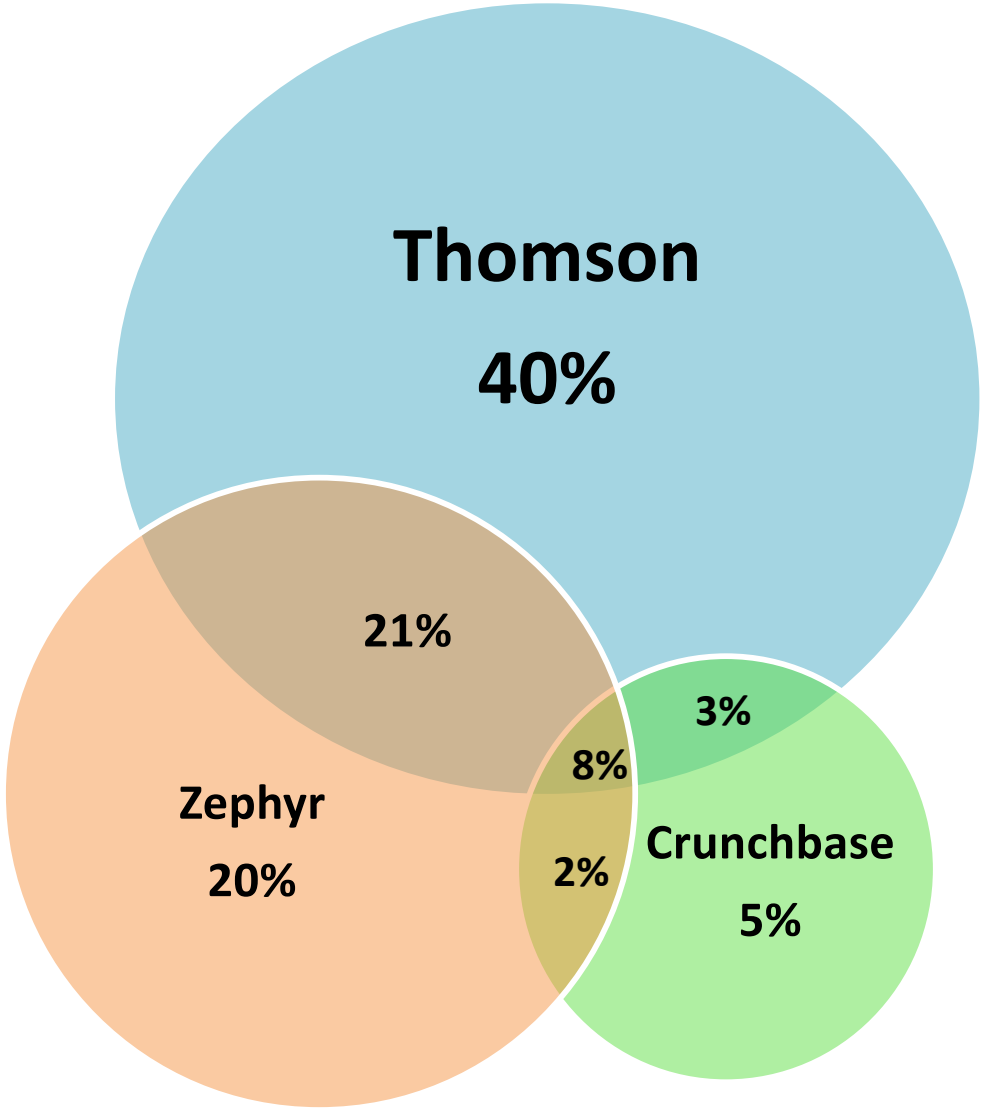
- Data are gathered from **3 source databases:**
  - **Thomson Eikon**
  - **BvD Zephyr**
  - **Crunchbase**

- We defined a unique list of companies to be included in VICO by merging information on companies that were recorded in the 3 source databases
  - Companies'/investors' names were disambiguated through fuzzy matching and manual checks

- Additional accounting information was collected from **BvD Orbis database**
  - Accounting data are available for 18,165 companies (75% of the sample) from **2005 to 2019**

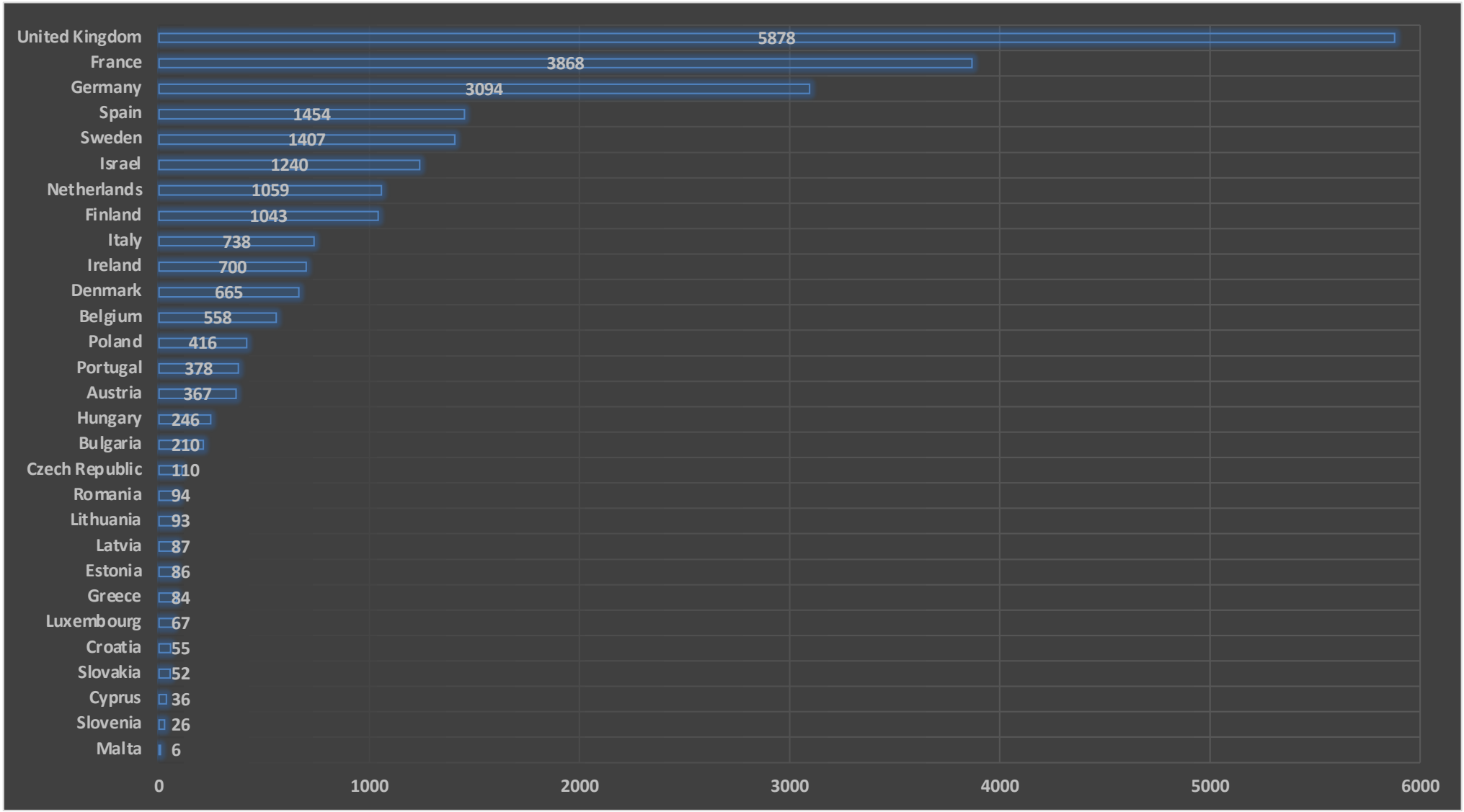
Companies distribution by data source

RISIS



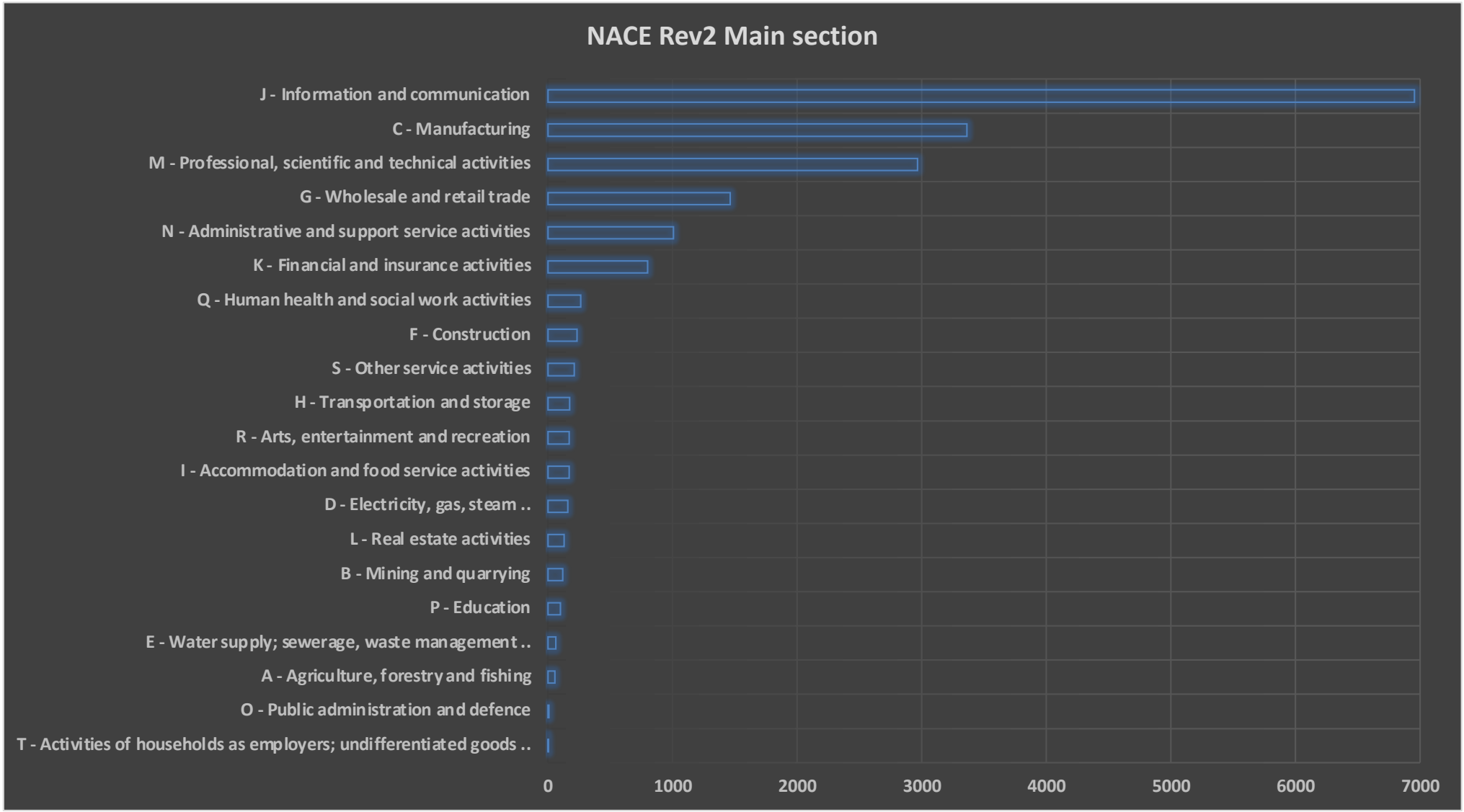
Company distribution by country

RISIS



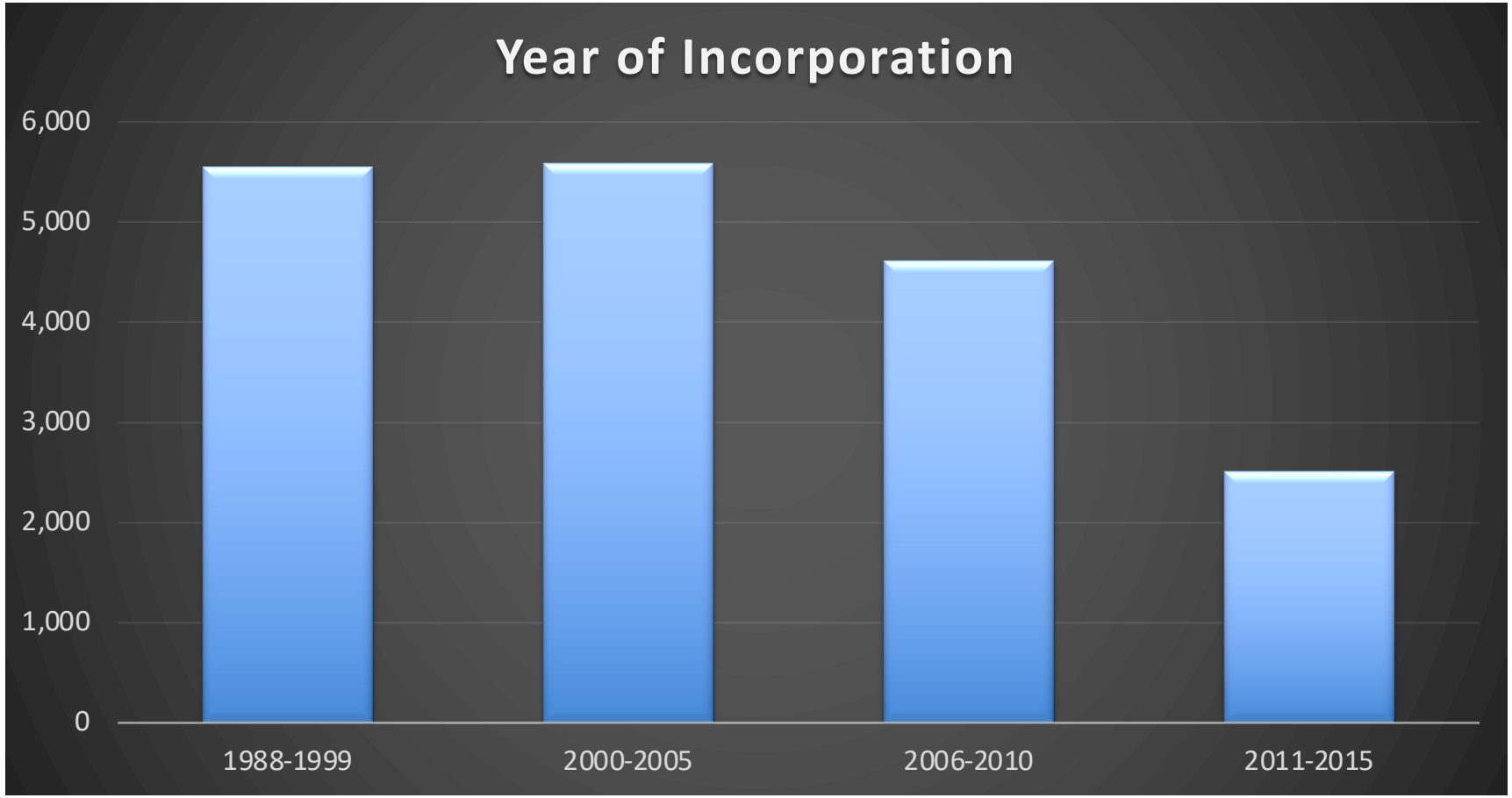
Company distribution by industry

RISIS



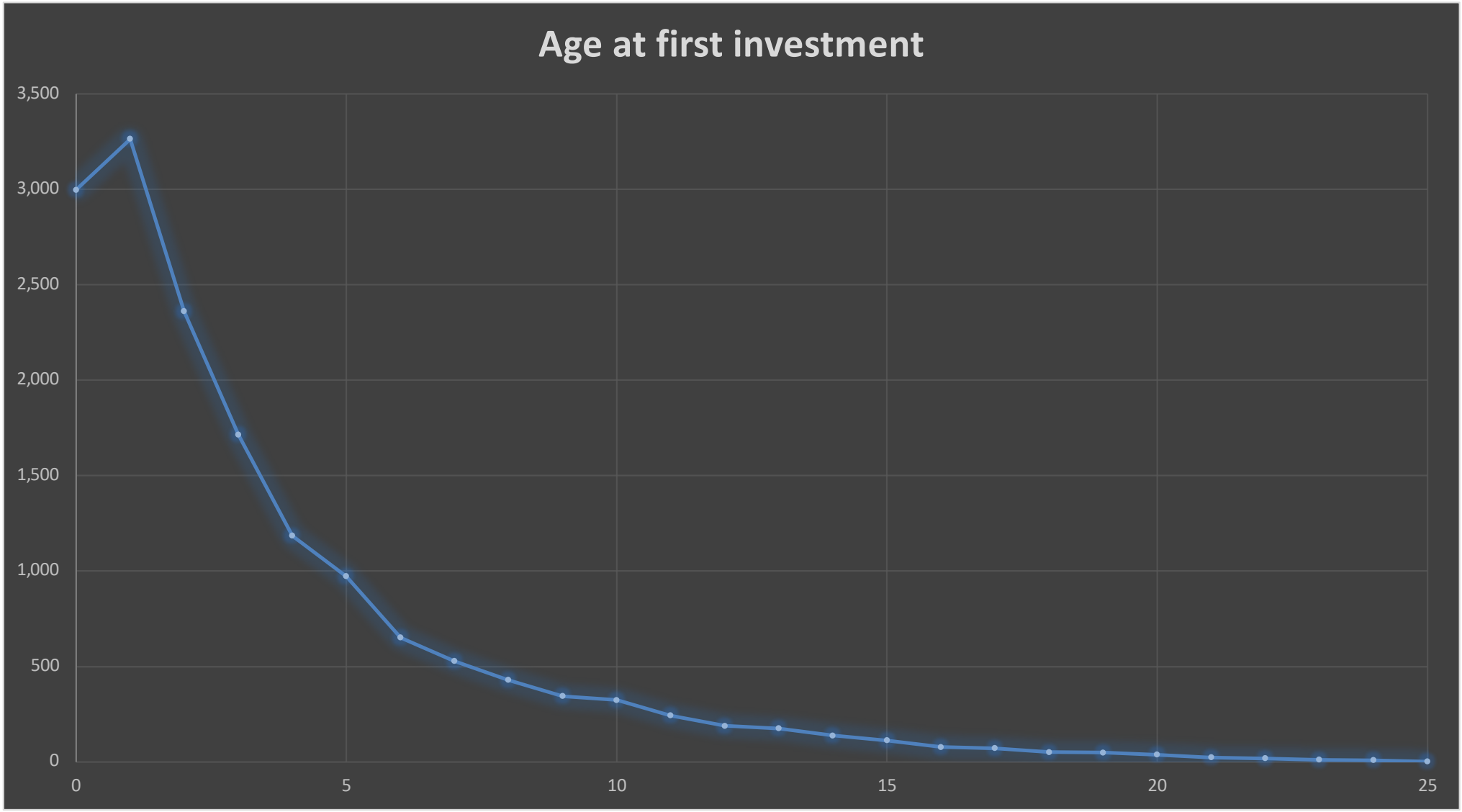
Company distribution by age

RISIS



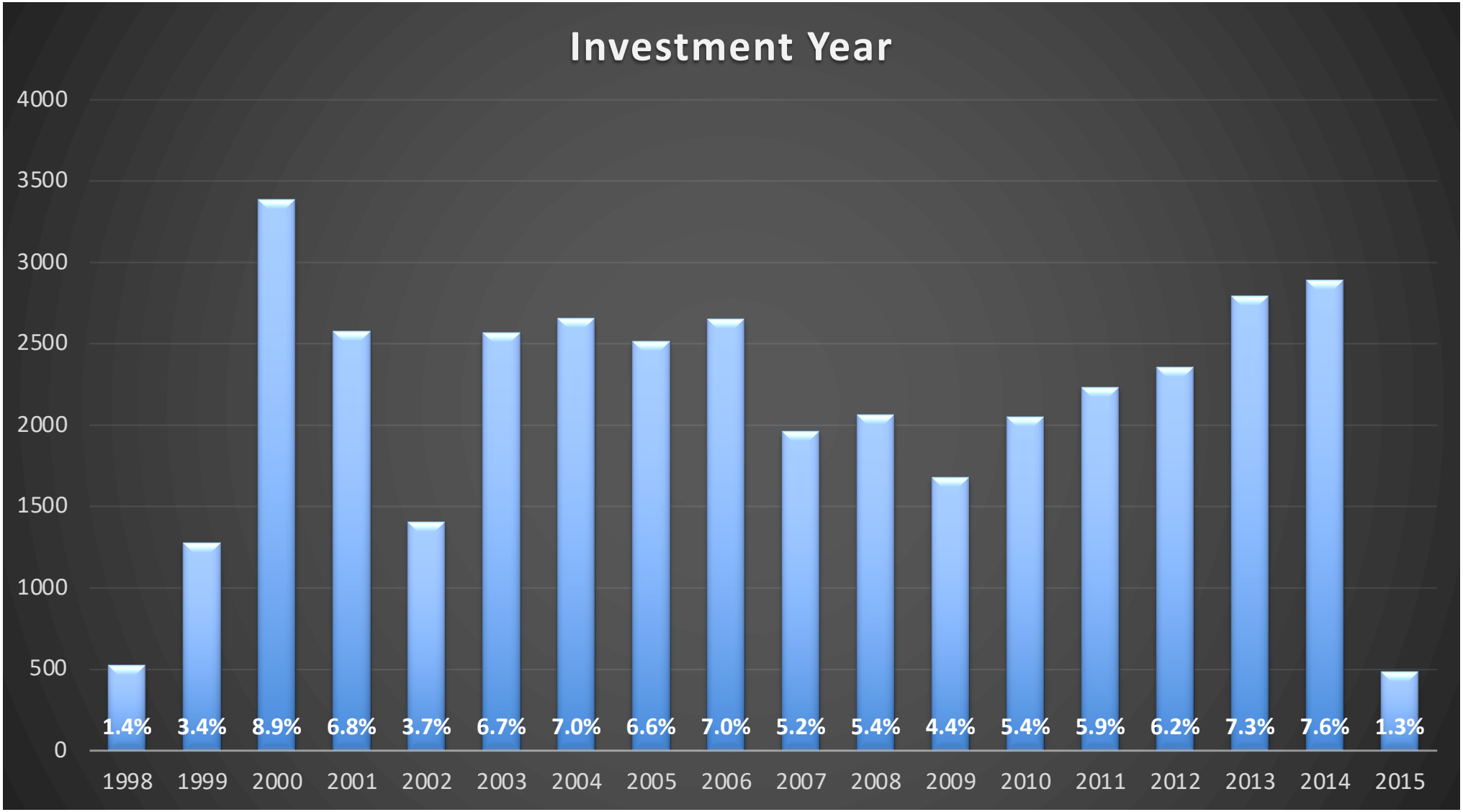
Company distribution by age at the time of first investment

RISIS



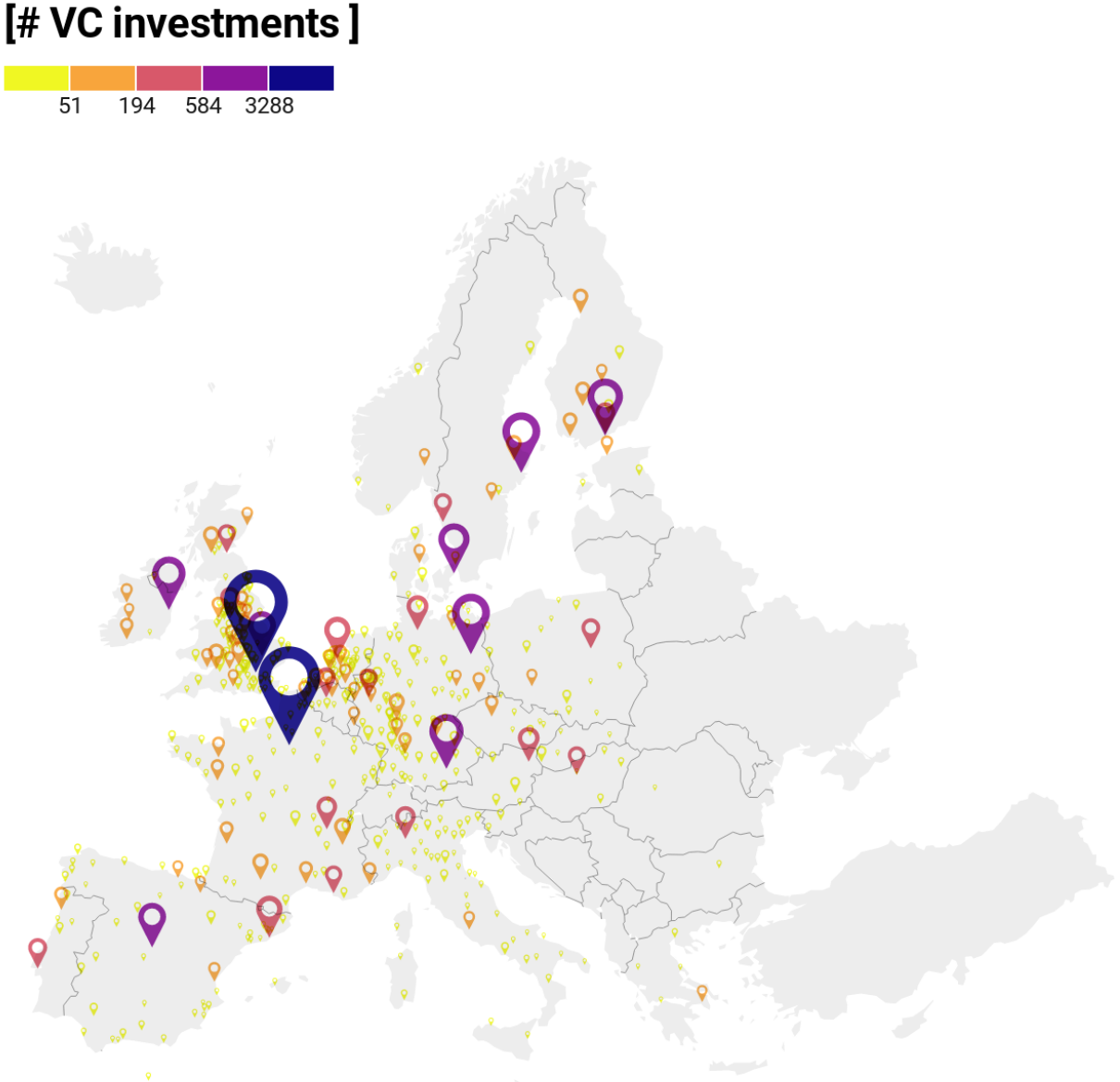
Investments

RISIS



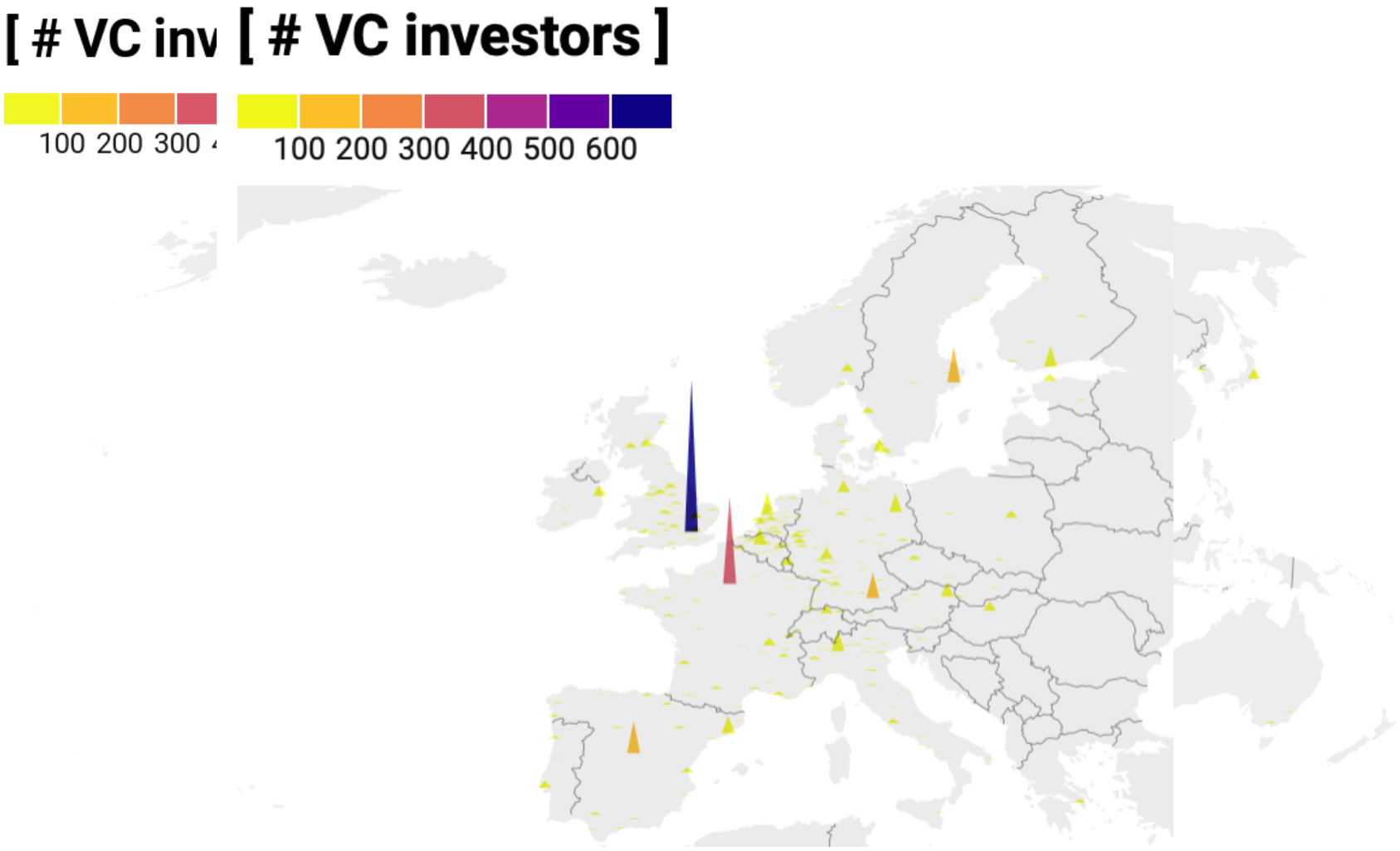
Investees' distribution by urban area

RISIS



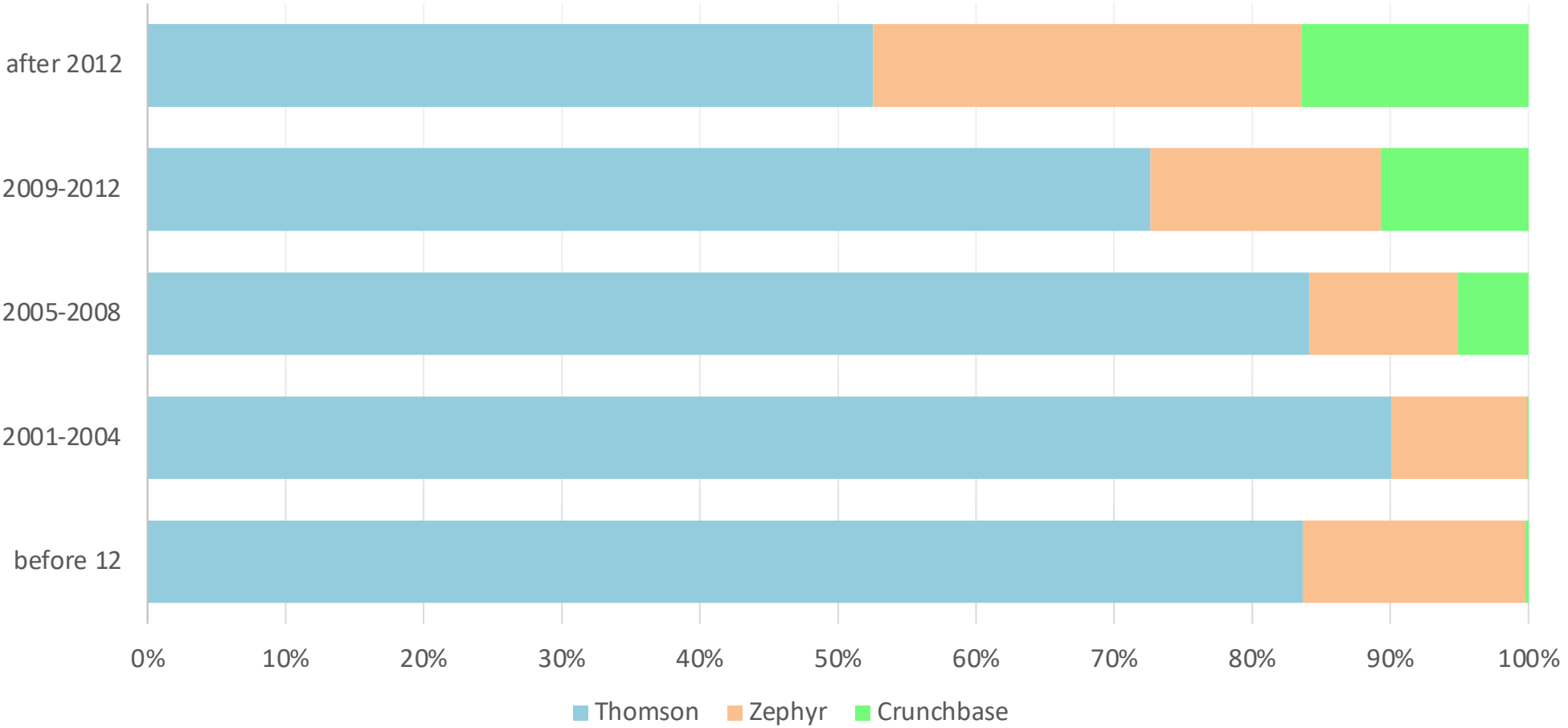
Investors distribution by urban area

RISIS



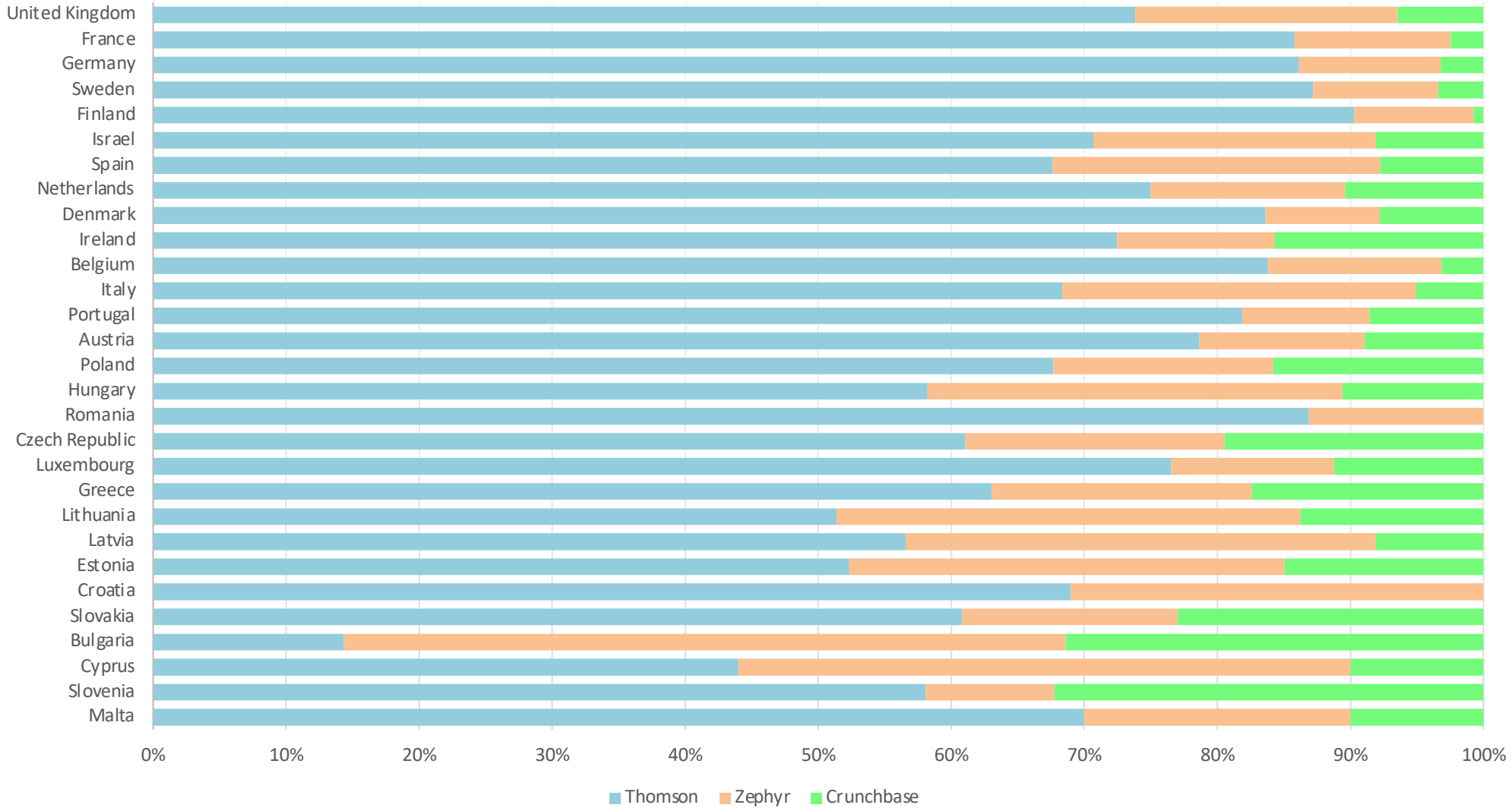
Investments by investment year and data source

RISIS



Investments by company's country and data source

RISIS



VICO

RISIS 

- To get access to the VICO dataset you can register to the RISIS Portal at this link and present a project proposal:

https://rcf.risis2.eu/datasets

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Panel Data in STATA

Aim of analysis

RISIS 

- The aim of our analysis is to study the **performance of VC-backed companies** after receiving VC investment
- We employ the VICO panel dataset
- Which performance are we interested in?
    - **Turnover/sales growth**
    - **Employees growth**
    - **Total assets growth**

Sample from VICO

RISIS

- We will use a **sample of data extracted from VICO:**
  - Companies receiving first VC investment within 10 years since incorporation
  - With available accounting information
  - Matched with a control sample of non-VC backed companies

- To build the control sample **1-to-1 Propensity Score Matching** was applied on a set of a priori firm's characteristics:
  - Country, region
  - Industry
  - Age at the time of first VC investment
  - Turnover at the time of first VC investment

- 2,850 VC backed + 2,850 non-VC backed firms

Sample from VICO

RISIS

Country	VC 0	1	Total
France	870	915	1,785
United Kingdom	418	393	811
Germany	326	314	640
Spain	306	312	618
Sweden	246	248	494
Finland	192	229	421
Italy	135	117	252
Bulgaria	47	46	93
Portugal	43	41	84
Hungary	37	33	70
Poland	34	31	65
Belgium	31	32	63
Latvia	25	22	47
Estonia	23	20	43
Romania	18	11	29
Austria	15	11	26
Denmark	13	13	26
Ireland	13	13	26
Netherlands	14	11	25
Czech Republic	10	12	22
Lithuania	9	7	16
Croatia	7	5	12
Slovakia	6	6	12
Luxembourg	5	3	8
Greece	2	2	4
Malta	3	1	4
Slovenia	2	2	4
Total	2,850	2,850	5,700

Industry	VC 0	1	Total
internet	986	1,002	1,988
medium-low tech	889	895	1,784
R&D and Engineering	344	357	701
high-tech manufactu..	269	246	515
Medical/Health/Life..	145	142	287
media	123	117	240
Biotechnology	94	91	185
Total	2,850	2,850	5,700

VC_hub_top 20F	VC 0	1	Total
0	1,343	1,322	2,665
1	1,507	1,528	3,035
Total	2,850	2,850	5,700

Sample from VICO

RISIS

ttest age, by(VC)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	2,850	1.369825	.0499348	2.665788	1.271913	1.467737
1	2,842	1.380014	.047641	2.539764	1.2866	1.473429
combined	5,692	1.374912	.0345071	2.603403	1.307265	1.442559
diff		-.0101895	.0690203		-.1454956	.1251165

diff = mean(0) - mean(1) t = -0.1476

Ho: diff = 0 degrees of freedom = 5690

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.4413 Pr(|T| > |t|) = 0.8826 Pr(T > t) = 0.5587

Econonometric specification RISIS

$$\text{Ln Growth}_{i,t} = a_0 + a_1 \text{Ln Size}_{i,t-1} + a_2 \text{Ln Age}_{i,t} + \mathbf{a_3 VC_{i,t-1}} + C_i + S_i + T_i + \alpha_i + u_{it}$$

$\text{Growth}_{i,t}$ is either sales, employees, assets growth

$\text{Size}_{i,t-1}$ is either sales, employees, assets in the prior period

$\text{VC}_{i,t-1}$ is a dummy variable that switch permanently from 0 to 1 in the year of receipt of the first round of VC

C_i are country dummies

S_i are industry dummies

T_i are year dummies

α_i is the unobserved entity-specific time-constant error term

u_{it} is the idiosyncratic error term

Panel data

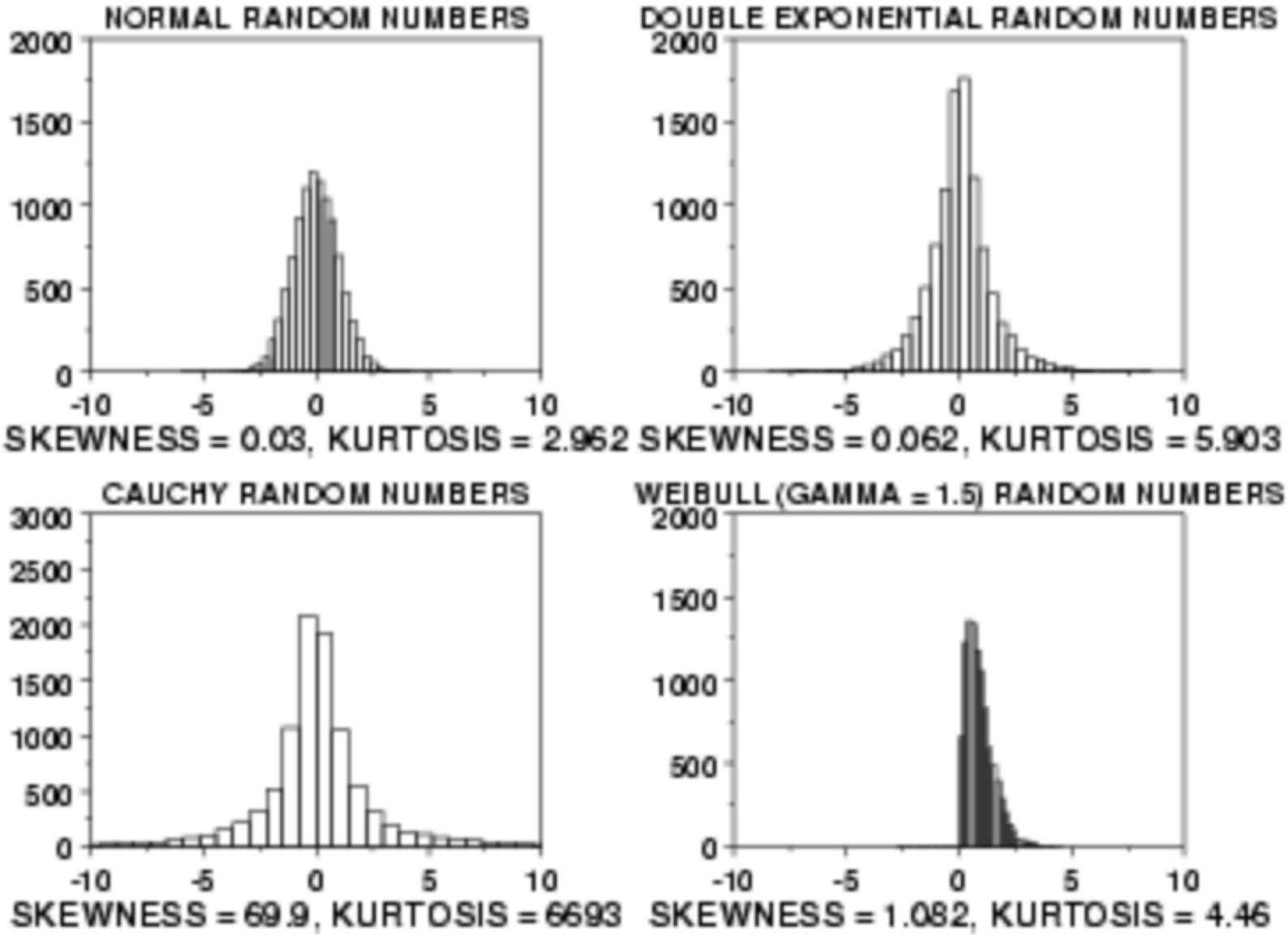
RISIS 

- We focus on three techniques to analyze panel data
  - **(Pooled OLS)**
  - **Fixed effects**
  - **Random effects**
  - **Instrumental variables regression**

- Before starting with any analysis, it is important to explore the data!

sum, hist

RISIS



Treatment of outliers

RISIS

- *Trimming* means discarding values at the tails of the distribution. That is, a percentage of the lowest and (an equal percentage of) the highest values of a variable are removed from the data
- *Winsorizing*: the values at the tails of the distribution are not removed, but are recoded to less extreme values
  - winsor *varname,* gen(*newvar*) h(#)
  - winsor *varname*, gen(*newvar*) p(#)

- Retain the outliers

xtsum

RISIS

[1.1] $\bar{x}_{..} \equiv \frac{1}{NT}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\right)$ (media *overall*);

[1.2] $\bar{x}_{i.} \equiv \frac{1}{N}\left[\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} x_{it}\right)\right]$ (media *between*);

[1.3] $\bar{x}_{.t} \equiv \frac{1}{T}\left[\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} x_{it}\right)\right]$ (media *within*);

[1.4] $\sqrt{\frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{x}_{..}\right)^2}$ (deviazione standard *overall*);

[1.5] $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_{i.}-\bar{x}_{i.}\right)^2}$ (deviazione standard *between*);

[1.6] $\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(x_{.t}-\bar{x}_{.t}\right)^2}$ (deviazione standard *within*).

xtsum – example

RISIS

id	score	$\bar{X}$	s
1	70	70	0
1	70		
2	70	80	14.14214
2	90		
3	90	60	42.42641
3	30		
Mean	70	70	18.85618

```
. xtsum score
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
score	overall	70	21.9089	30	90	N =	6
	between		10	60	80	n =	3
	within		20	40	100	T =	2

xtsum – example

RISIS

i	x_{it}	$\bar{x}_i$	$x_{it} - \bar{x}_i$	$x_{it} - \bar{x}_i + \bar{\bar{x}}$
1	70	70	0	70
1	70		0	70
2	70	80	-10	60
2	90		10	80
3	90	60	30	100
3	30		-30	40
Mean	70	70	0	70

```
. xtsum score
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
score	overall	70	21.9089	30	90	N =	6
	between		10	60	80	n =	3
	within		20	40	100	T =	2

Fixed Effect (FE) model

RISIS

- Standard equation for FE model is as follows:

$$Y_{it} = \beta_1 X_{it} + \underbrace{\alpha_i + u_{it}}_{\varepsilon_{it}}$$

- Where:
  - Y_{it} is the dependent variable
  - X_{it} is the independent variable
  - β_1 is the coefficient for the independent variable
  - α_i is the unobserved entity-specific time-constant error term. **It can be correlated with X_{it}**
  - u_{it} is the idiosyncratic error term that varies across individuals and time. It is assumed to be uncorrelated with X_{it}

Fixed Effect (FE) model

RISIS

- Ideally, if we could include in the econometric specification a dummy variable for each entity the unobserved entity-specific heterogeneity would be controlled for with a simple OLS regression

- The equation for the FE model would become (**Least Squares Dummy Variables LSDV - estimator**):

$$Y_{it} = \beta_1 X_{it} + \alpha_1 + \alpha_2 D_2 + \cdots + \alpha_N D_N + u_{it}$$

- However, when T is small and N is large this model cannot be estimated

Fixed Effect (FE) model

RISIS

- **A solution is to eliminate the fixed characteristics α_i** through mean-differencing (**Within Group estimator**)
  - Model to be estimated: $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$
  - Entity specific means over t: $\overline{Y}_i = \beta_1 \overline{X}_i + \alpha_i + \overline{u}_{it}$

- Within transformation ('demeaning' the data):

$$Y_{it} - \overline{Y}_i = \beta_1 (X_{it} - \overline{X}_i) + (u_{it} - \overline{u}_{it})$$

- This way **FE models control for all time invariant differences between the entities (by eliminating α_i)** , so the estimated coefficients are not biased because of unobserved time-invariant heterogeneity (*xtreg, fe* in STATA)

Fixed effects – STATA output

RISIS

$$Y_{it} = \beta_1 X_{it} + \ldots + \beta_k X_{kt} + \alpha_i + e_{it} \quad \text{[see eq.1]}$$

Outcome variable | Predictor variable(s)

Fixed effects option

Total number of cases (rows)

Total number of groups (entities)

```
. xtreg  y  x1, fe

Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

The errors u_i are correlated with the regressors in the fixed effects model

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

Coefficients of the regressors. Indicate how much *Y* changes when *X* increases by one unit.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

29.7% of the variance is due to differences across panels.

'rho' is known as the intraclass correlation

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

sigma_u = sd of residuals within groups u_i

sigma_e = sd of residuals (overall error term) e_i

Random Effect (RE) model

RISIS

- Standard equation for RE model is as follows:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \underbrace{\alpha_i + u_{it}}_{\varepsilon_{it}}$$

- Where:
  - Y_{it} is the dependent variable
  - X_{it} is the independent variable
  - β_1 is the coefficient for the independent variable
  - β_0 is the intercept
  - ε_{it} is the error term, disaggregated in the two components α_i and u_{it}

- **Key assumption: α_i are i.i.d. random-effects not correlated to X_{it}**

Random Effect (RE) model

RISIS

- Standard equation for RE model is as follows:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \underbrace{\alpha_i + u_{it}}_{\varepsilon_{it}}$$

- $E(\varepsilon_{it}) = 0$
- $Var(\varepsilon_{it}) = Var(\alpha_i) + Var(u_{it})$, (sum of within and between component variances)
- If $Var(\alpha_i) = 0$ then $Var(\varepsilon_{it}) = Var(u_{it})$, and there is no difference between the pooled regression model and the RE model

→ Breusch-Pagan test

Random effects – STATA output

RISIS

Outcome variable | Predictor variable(s) | Random effects option

```
. xtreg y x1, re

Random-effects GLS regression                   Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

Random effects u_i ~ Gaussian                   Wald chi2(1)       =      1.91
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.1669

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.25e+09   9.02e+08     1.38   0.167    -5.21e+08    3.02e+09
       _cons |   1.04e+09   7.91e+08     1.31   0.190    -5.13e+08    2.59e+09
-------------+----------------------------------------------------------------
     sigma_u |  1.065e+09
     sigma_e |  2.796e+09
         rho |  .12664193   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Differences across units are uncorrelated with the regressors

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Interpretation of the coefficients is tricky since they include both the within-entity and between-entity effects. In the case of TSCS data represents the average effect of *X* over *Y* when *X* changes across time and between countries by one unit.

Statistical tests sum-up

RISIS

FE vs. OLS	RE vs. OLS	Your Model
$H_0 = \mu_1 = \mu_2 = \ldots = \mu$ F or Wald Test	$H_0 = \mathrm{Var}(\mu_i) = 0$ Breusch-Pagan Test	
H_0 not rejected $\Rightarrow$ No FE	H_0 not rejected $\Rightarrow$ No RE	Pooled OLS
H_0 rejected $\Rightarrow$ FE	H_0 not rejected $\Rightarrow$ No RE	FE Model
H_0 not rejected $\Rightarrow$ No FE	H_0 rejected $\Rightarrow$ RE	RE Model
H_0 rejected $\Rightarrow$ FE	H_0 rejected $\Rightarrow$ RE	Choose one based on Hausman test.