The Role of Public Statistics in the XXIst Century Data Revolution

Frederic Udina

Institut d'Estadística de Catalunya

Jornades de consultoria estadística i software II

Centre de Recerca Matemàtica, 24 d'octubre del 2013

Generalitat de Catalunya
Institut d'Estadística
de Catalunya

STATISTICS2013.ORG
Any Internacional de
l'ESTADÍSTICA
Organisme participant

Esquema

Statistics in IYS2013

Data revolution

A system of statistical registers

Statistics in IYS2013

- Origin, two big spread moments

Data revolution

- Data, data, data everywhere

A system of statistical registers

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Origins

- In Catalonia, there were *fogatges* in XIVth century: counting of dwellings!
- Even older census, e.g. in the Bible.
- Census aimed to impose taxes, or to recruit armies. Now are for planning social policies.
- *Statistics* comes from *state*, official statistics is in the origins.

Origins

- In Catalonia, there were *fogatges* in XIVth century: counting of dwellings!
- Even older census, e.g. in the Bible.
- Census aimed to impose taxes, or to recruit armies. Now are for planning social policies.
- *Statistics* comes from *state*, official statistics is in the origins.

Origins

- In Catalonia, there were *fogatges* in XIVth century: counting of dwellings!
- Even older census, e.g. in the Bible.
- Census aimed to impose taxes, or to recruit armies. Now are for planning social policies.
- *Statistics* comes from *state*, official statistics is in the origins.

Origins

- In Catalonia, there were *fogatges* in XIVth century: counting of dwellings!
- Even older census, e.g. in the Bible.
- Census aimed to impose taxes, or to recruit armies. Now are for planning social policies.
- *Statistics* comes from *state*, official statistics is in the origins.

Origins: first Bills of Mortality

- Analized by John Graunt, XVIIth century: origin of demography.
- In XVIII-XIX centuries many examples of collecting and analyzing data: mainly for public statistics.
- John Snow and the colera epidemy, Londres 1854: getting evidence from data.
- Anyway, statistics is a very young science!

Origins: first Bills of Mortality

- Analized by John Graunt, XVIIth century: origin of demography.
- In XVIII-XIX centuries many examples of collecting and analyzing data: mainly for public statistics.
- John Snow and the colera epidemy, Londres 1854: getting evidence from data.
- Anyway, statistics is a very young science!

Origins: first Bills of Mortality

- Analized by John Graunt, XVIIth century: origin of demography.
- In XVIII-XIX centuries many examples of collecting and analyzing data: mainly for public statistics.
- John Snow and the colera epidemy, Londres 1854: getting evidence from data.
- Anyway, statistics is a very young science!

Origins: first Bills of Mortality

- Analized by John Graunt, XVIIth century: origin of demography.
- In XVIII-XIX centuries many examples of collecting and analyzing data: mainly for public statistics.
- John Snow and the colera epidemy, Londres 1854: getting evidence from data.
- Anyway, statistics is a very young science!

Then probability came in

- Somewhat incredible, the science of exactness dealing with randomness!
- Bernoulli, De Moivre, Laplace, Gauss. . .

That young science moves fast.

When data met probability the first big spread of statistics appear. Statistics entered in virtually all fields of science.

Note that the Higgs boson only exists statistically!

One could say

data + model → Statistics

Then probability came in

- Somewhat incredible, the science of exactness dealing with randomness!
- Bernoulli, De Moivre, Laplace, Gauss. . .

That young science moves fast.

When data met probability the first big spread of statistics appear.
Statistics entered in virtually all fields of science.

Note that the Higgs boson only exists statistically!

One could say

data + model $\rightarrow$ Statistics

Then probability came in

- Somewhat incredible, the science of exactness dealing with randomness!
- Bernoulli, De Moivre, Laplace, Gauss. . .

That young science moves fast.

When data met probability the first big spread of statistics appear.

Statistics entered in virtually all fields of science.

Note that the Higgs boson only exists statistically!

One could say

data + model → Statistics

Then probability came in

- Somewhat incredible, the science of exactness dealing with randomness!
- Bernoulli, De Moivre, Laplace, Gauss. . .

That young science moves fast.

When data met probability the first big spread of statistics appear.
Statistics entered in virtually all fields of science.

Note that the Higgs boson only exists statistically!

One could say

data + model → Statistics

Then probability came in

- Somewhat incredible, the science of exactness dealing with randomness!
- Bernoulli, De Moivre, Laplace, Gauss. . .

That young science moves fast.

When data met probability the first big spread of statistics appear.

Statistics entered in virtually all fields of science.

Note that the Higgs boson only exists statistically!

One could say

data + model → Statistics

Statistics in IYS2013

Origin, two big spread moments

Data revolution

Data, data, data everywhere

A system of statistical registers

A new paradigm

With XXIst century we get Big Data.
What is it? Is it a real change?
May it be that now

data + algorithms → Statistics

An interesting BBC video

Horizon - Age of Big Data.BBC.2013.Data Mining - Bons Legenda...

BBC video

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Video contents

- Trying to predict next crime location in L.A. area
- Predicting financial markets
- Personalized advertising
- Monitoring illness using personal data: a case of atrial fibrillation

How big is big data? Is it only a matter of size?

Is it compatible with offical statistics?

Video contents

- Trying to predict next crime location in L.A. area
- Predicting financial markets
- Personalized advertising
- Monitoring illness using personal data: a case of atrial fibrillation

How big is big data? Is it only a matter of size?

Is it compatible with offical statistics?

Video contents

- Trying to predict next crime location in L.A. area
- Predicting financial markets
- Personalized advertising
- Monitoring illness using personal data: a case of atrial fibrillation

How big is big data? Is it only a matter of size?

Is it compatible with offical statistics?

How to mix oil and water

Big data is

- Fast to produce, to collect
- Volatile, it changes fast
- Privacy limits are fuzzy
- Mainly owned by private corp.

Official data is

- More slow to produce, compulsory to provide
- It needs to be solid and safe
- Privacy limits well defined
- Owned and managed publicly

But there are many ways to cooperate!

- For research
- Commercially. . .
- With other administrations

How to mix oil and water

Big data is

- Fast to produce, to collect
- Volatile, it changes fast
- Privacy limits are fuzzy
- Mainly owned by private corp.

Official data is

- More slow to produce, compulsory to provide
- It needs to be solid and safe
- Privacy limits well defined
- Owned and managed publicly

But there are many ways to cooperate!

- For research
- Commercially. . .
- With other administrations

How to mix oil and water

Big data is

- Fast to produce, to collect
- Volatile, it changes fast
- Privacy limits are fuzzy
- Mainly owned by private corp.

Official data is

- More slow to produce, compulsory to provide
- It needs to be solid and safe
- Privacy limits well defined
- Owned and managed publicly

But there are many ways to cooperate!

- For research
- Commercially. . .
- With other administrations

How to mix oil and water

Big data is

- Fast to produce, to collect
- Volatile, it changes fast
- Privacy limits are fuzzy
- Mainly owned by private corp.

Official data is

- More slow to produce, compulsory to provide
- It needs to be solid and safe
- Privacy limits well defined
- Owned and managed publicly

But there are many ways to cooperate!

- For research
- Commercially. . .
- With other administrations

Statistics in IYS2013

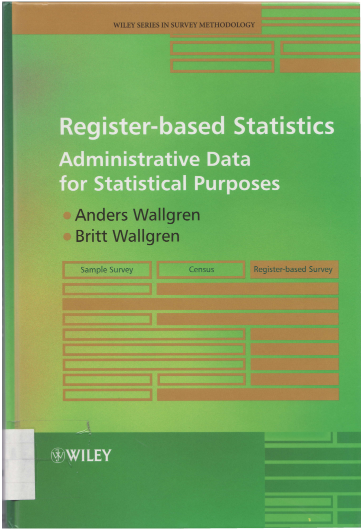
Origin, two big spread moments

Data revolution

Data, data, data everywhere

A system of statistical registers

A key reference



WILEY SERIES IN SURVEY METHODOLOGY

Register-based Statistics

Administrative Data for Statistical Purposes

- Anders Wallgren
- Britt Wallgren

“Although this is the oldest and most common form of statistics, no well stablished theory in the field exists”

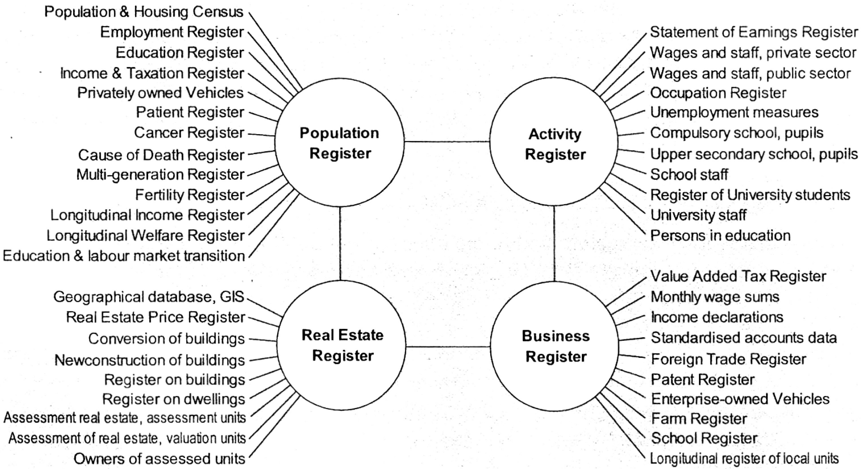
[. . .]

“One important reason of this shortfall is that the subject field of register-based surveys is not included in academic statistics” [. . .]

“Unfortunately statistical science has so far not included any theory on statistical systems.”

A model of system of statistical registers

Chart 2.10 A system of statistical registers – registers by object type and subject field



Population Register

Population & Housing Census
Employment Register
Education Register
Income & Taxation Register
Privately owned Vehicles
Patient Register
Cancer Register
Cause of Death Register
Multi-generation Register
Fertility Register
Longitudinal Income Register
Longitudinal Welfare Register
Education & labour market transition

Activity Register

Statement of Earnings Register
Wages and staff, private sector
Wages and staff, public sector
Occupation Register
Unemployment measures
Compulsory school, pupils
Upper secondary school, pupils
School staff
Register of University students
University staff
Persons in education

Real Estate Register

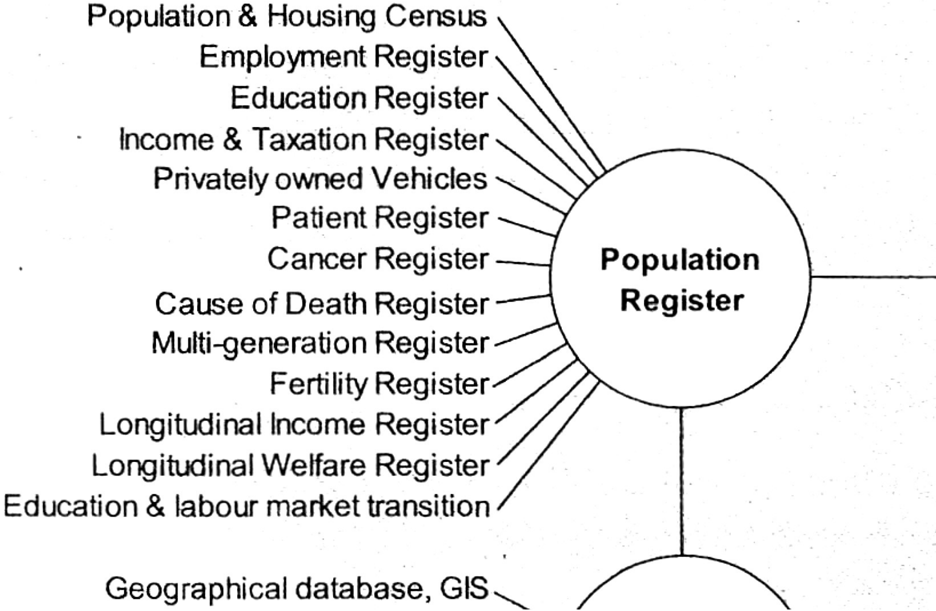
Geographical database, GIS
Real Estate Price Register
Conversion of buildings
Newconstruction of buildings
Register on buildings
Register on dwellings
Assessment real estate, assessment units
Assessment of real estate, valuation units
Owners of assessed units

Business Register

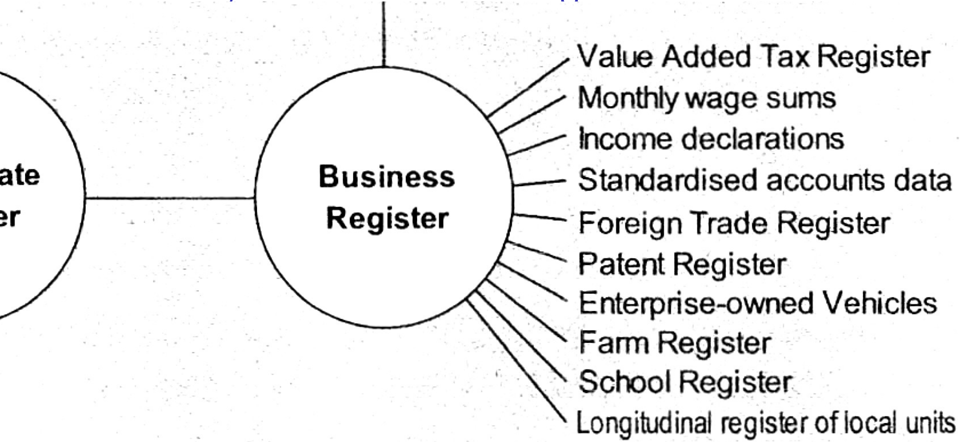
Value Added Tax Register
Monthly wage sums
Income declarations
Standardised accounts data
Foreign Trade Register
Patent Register
Enterprise-owned Vehicles
Farm Register
School Register
Longitudinal register of local units

from Wallgren's book

A model of system of statistical registers

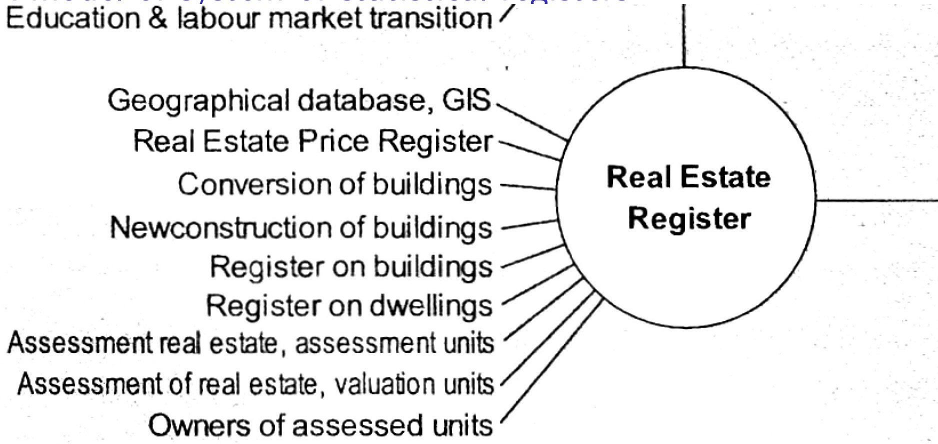

Population & Housing Census
Employment Register
Education Register
Income & Taxation Register
Privately owned Vehicles
Patient Register
Cancer Register
Cause of Death Register
Multi-generation Register
Fertility Register
Longitudinal Income Register
Longitudinal Welfare Register
Education & labour market transition
Geographical database, GIS
Population Register

A model of system of statistical registers



from Wallgren's bo

A model of system of statistical registers



Education & labour market transition

Geographical database, GIS

Real Estate Price Register

Conversion of buildings

Newconstruction of buildings

Register on buildings

Register on dwellings

Assessment real estate, assessment units

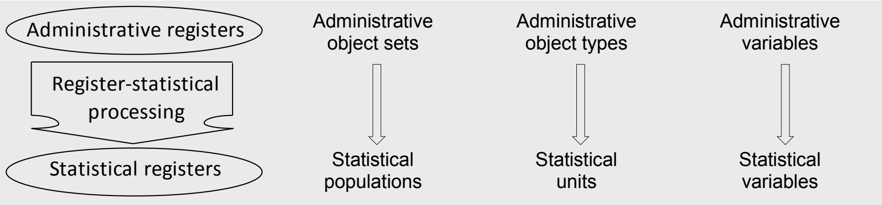
Assessment of real estate, valuation units

Owners of assessed units

Real Estate Register

Transforming registers

Chart 1.6a From administrative registers to statistical registers



Administrative registers

Register-statistical processing

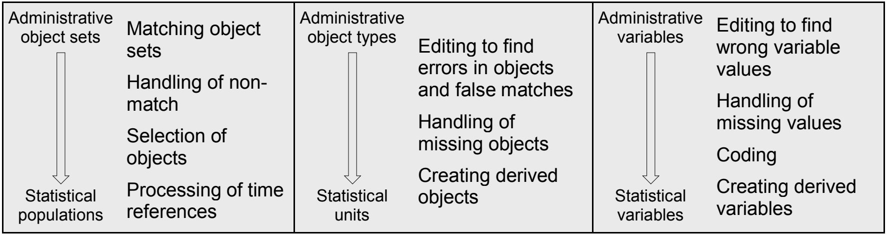
Statistical registers

Administrative object sets → Statistical populations

Administrative object types → Statistical units

Administrative variables → Statistical variables

Chart 1.6b From administrative registers to statistical registers



Administrative object sets → Statistical populations

Matching object sets

Handling of non-match

Selection of objects

Processing of time references

Administrative object types → Statistical units

Editing to find errors in objects and false matches

Handling of missing objects

Creating derived objects

Administrative variables → Statistical variables

Editing to find wrong variable values

Handling of missing values

Coding

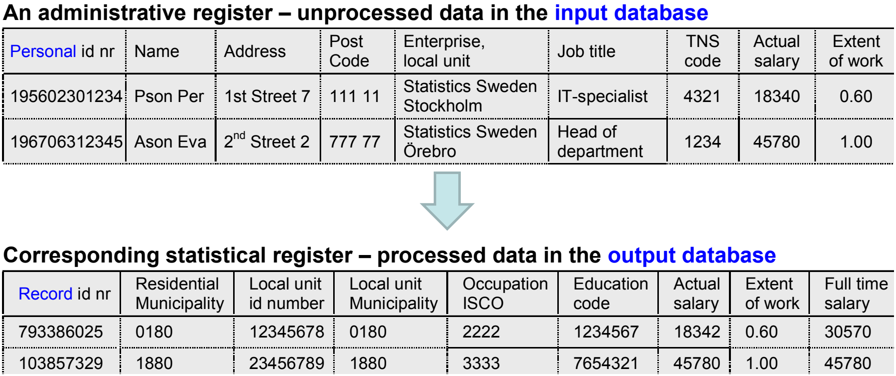
Creating derived variables

From Wallgren's course

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Protecting privacy and confidentiality

The input database is highly protected, only a few people can access it.



An administrative register – unprocessed data in the input database

Personal id nr	Name	Address	Post Code	Enterprise, local unit	Job title	TNS code	Actual salary	Extent of work
195602301234	Pson Per	1st Street 7	111 11	Statistics Sweden Stockholm	IT-specialist	4321	18340	0.60
196706312345	Ason Eva	2nd Street 2	777 77	Statistics Sweden Örebro	Head of department	1234	45780	1.00

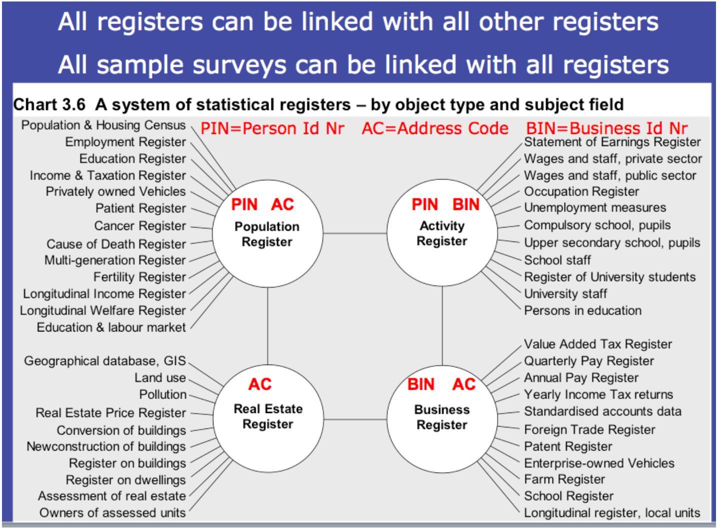
Corresponding statistical register – processed data in the output database

Record id nr	Residential Municipality	Local unit id number	Local unit Municipality	Occupation ISCO	Education code	Actual salary	Extent of work	Full time salary
793386025	0180	12345678	0180	2222	1234567	18342	0.60	30570
103857329	1880	23456789	1880	3333	7654321	45780	1.00	45780

The output database may be used by statisticians, subject to an agreement of confidentiality.

From Wallgren's course

Registers are *dissociated* but fully linked



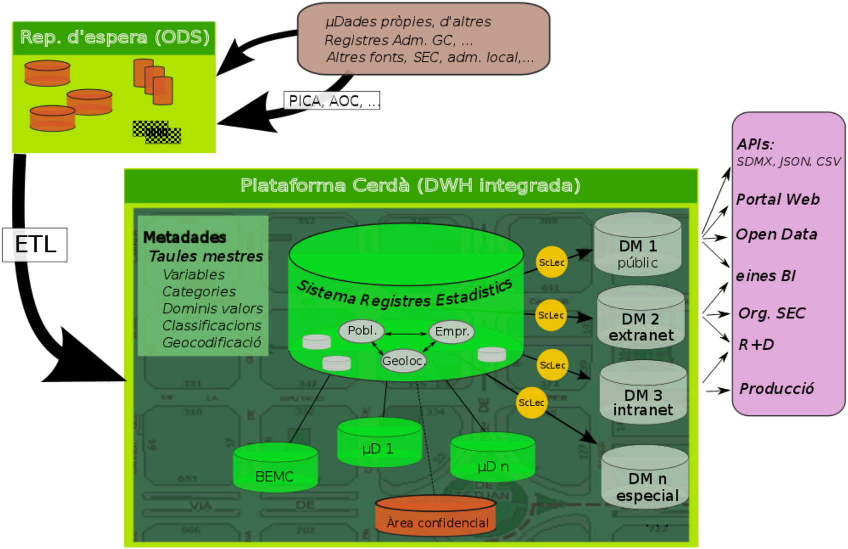
All registers can be linked with all other registers

All sample surveys can be linked with all registers

Chart 3.6 A system of statistical registers – by object type and subject field

From Wallgren's course

Organizing information in an integrated platform



ODS: Operational Data Storage ETL: Extract/Transform/Load DWH: Data warehouse DM: Data Mart BI: Bussiness Inteligence

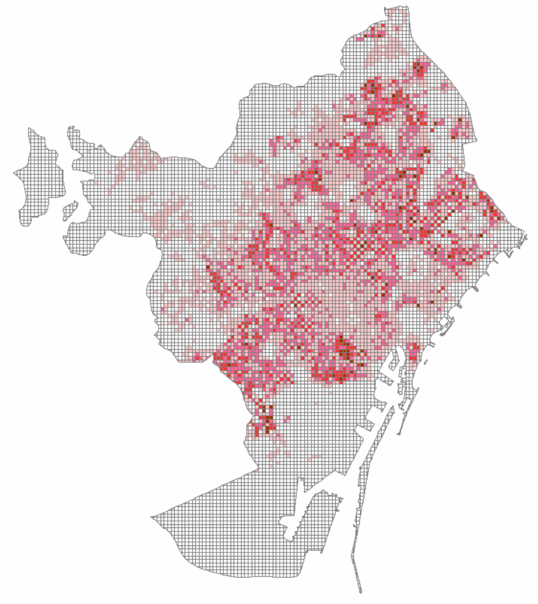
Generalitat de Catalunya
Institut d'Estadística de Catalunya

Preconditions for Register-based Statistics

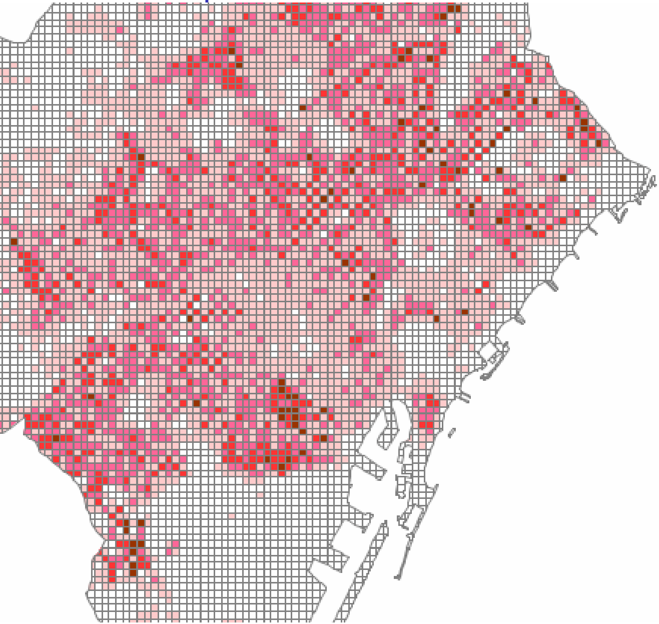
- Legal base for the NSO to use administrative data in a way that protects integrity
- Public approval for using administrative data for statistical purposes
- Unified systems of identity codes used in all sources
- Reliable administrative systems
- Cooperation between administrative authorities
- Regional and/or National Administrative Registers?

From Wallgren's course

Some examples



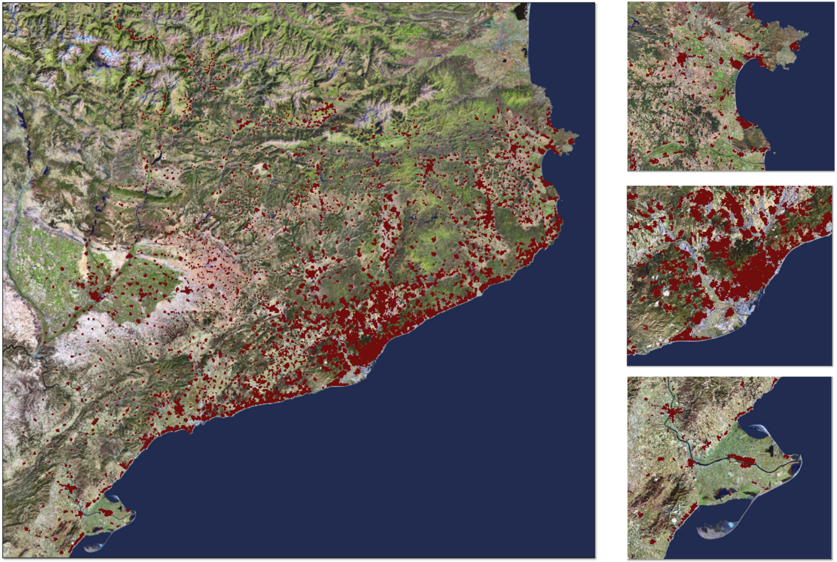
Some examples



Generalitat de Catalunya
Institut d'Estadística
de Catalunya

Some examples

Població de Catalunya a 1 de gener de 2013



Font: Idescat. ICC.

Generalitat de Catalunya
Institut d'Estadística de Catalunya