DAT4CTION

Torino, 19 novembre 2015

Estimating the diffusion of energy sources by analyzing web searches

An applicative study

Francesco Tarasconi
Vittorio Di Tomaso

csi piemonte

TORINO METROPOLI
Città metropolitana di Torino

CELI
LANGUAGE TECHNOLOGY

Problem and methodology

We would like to introduce…

> ”
>
> A methodology to estimate
> the usage of energy sources
> by assessing information need
> from web searches

Motivation

> In some cases, solar thermal in particular, we are missing exact (and ***georeferenced***) data to determine the usage by the population

The idea

> Use information needs, identified from web searches, to estimate usage of a specific source

Goals

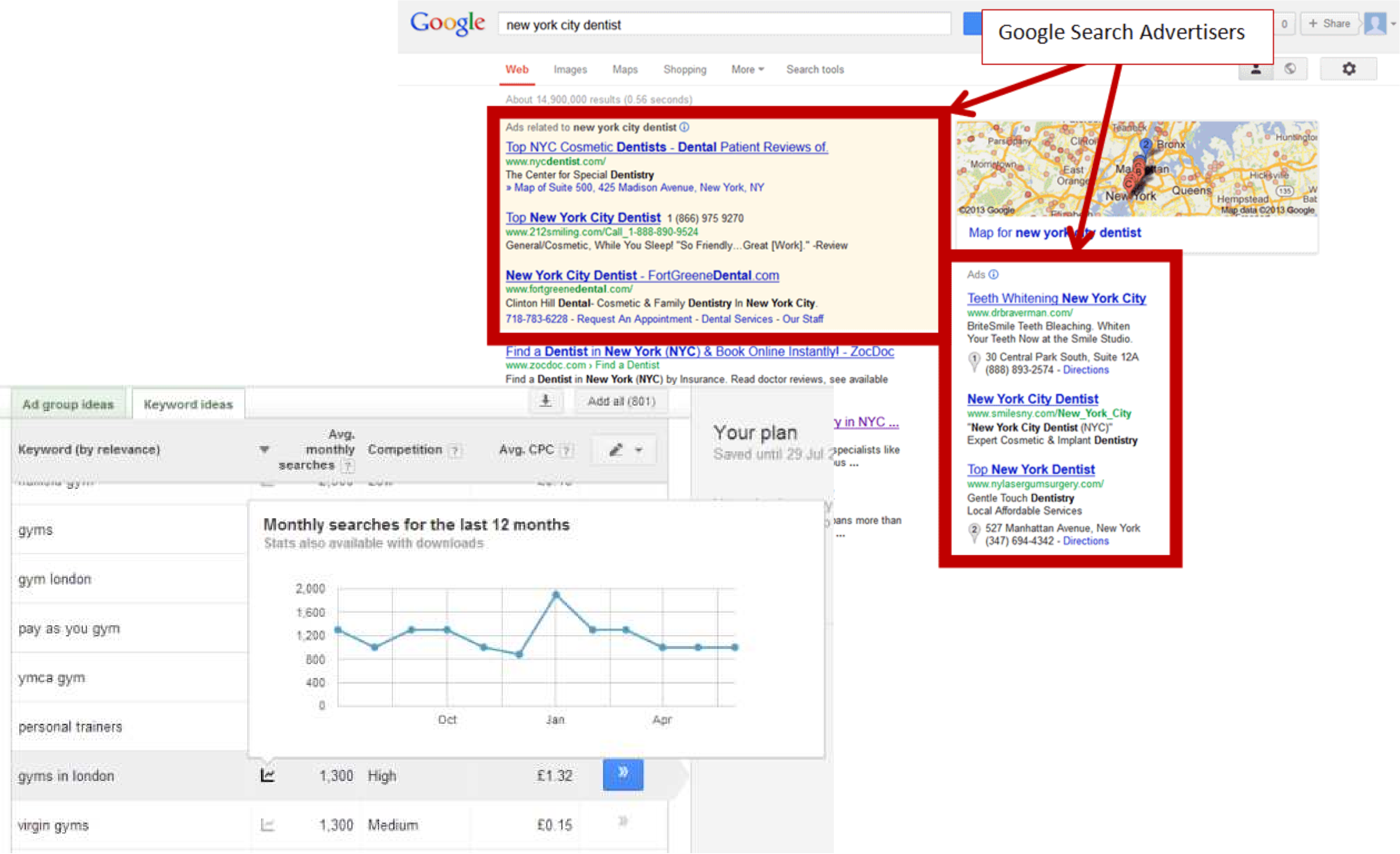
1. To study the usage of an energy source among the population using georeferenced online search volumes on related keywords, through Google and partners.
2. To estimate how an aggregated national value is distributed on the territory: Region by Region, Province by Province, …

Method

1. Gather data on cases with known usage:
   - methane fueled cars;
   - photovoltaic installations.
2. Extract lists of keywords related to the subject and frequently searched on the web.
3. Review lists (manually and using natural language technology tools) to better capture phenomena of interest.
4. Establish correlations between search volumes and usage.
5. Learn a predictive model.
6. Compare with other predictors: have we learned something new?
7. Generalize to other cases:
   - solar thermal systems.

Background

Google Search and AdWords



Google Search Advertisers

About 14,900,000 results (0.56 seconds)

Ads related to **new york city dentist**

Top NYC Cosmetic **Dentists** - **Dental** Patient Reviews of.
www.nyc**dentist**.com/
The Center for Special **Dentistry**
» Map of Suite 500, 425 Madison Avenue, New York, NY

Top **New York City Dentist** 1 (866) 975 9270
www.212smiling.com/Call_1-888-890-9524
General/Cosmetic, While You Sleep! "So Friendly...Great [Work]." -Review

New York City Dentist - FortGreene**Dental**.com
www.fortgreene**dental**.com/
Clinton Hill **Dental**- Cosmetic & Family **Dentistry** In **New York City**.
718-783-6228 - Request An Appointment - Dental Services - Our Staff

Find a **Dentist** in **New York (NYC)** & Book Online Instantly! - ZocDoc
www.zocdoc.com › Find a Dentist
Find a **Dentist** in **New York** (**NYC**) by Insurance. Read doctor reviews, see available

Map for **new york city dentist**

Ads

Teeth Whitening **New York City**
www.drbraverman.com/
BriteSmile Teeth Bleaching. Whiten Your Teeth Now at the Smile Studio.
30 Central Park South, Suite 12A
(888) 893-2574 - Directions

New York City Dentist
www.smilesny.com/New_York_City
"**New York City Dentist** (NYC)"
Expert Cosmetic & Implant **Dentistry**

Top **New York Dentist**
www.nylasergumsurgery.com/
Gentle Touch **Dentistry**
Local Affordable Services
527 Manhattan Avenue, New York
(347) 694-4342 - Directions

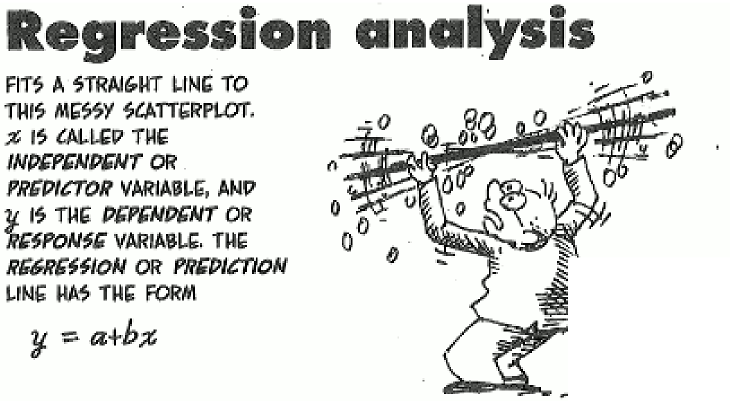
Keyword (by relevance)		Avg. monthly searches	Competition	Avg. CPC	
gyms					
gym london					
pay as you gym					
ymca gym					
personal trainers					
gyms in london		1,300	High	£1.32	»
virgin gyms		1,300	Medium	£0.15	»

Monthly searches for the last 12 months

Stats also available with downloads

Copyright © CELI s.r.l.

A Class of Predictive Models: Regression

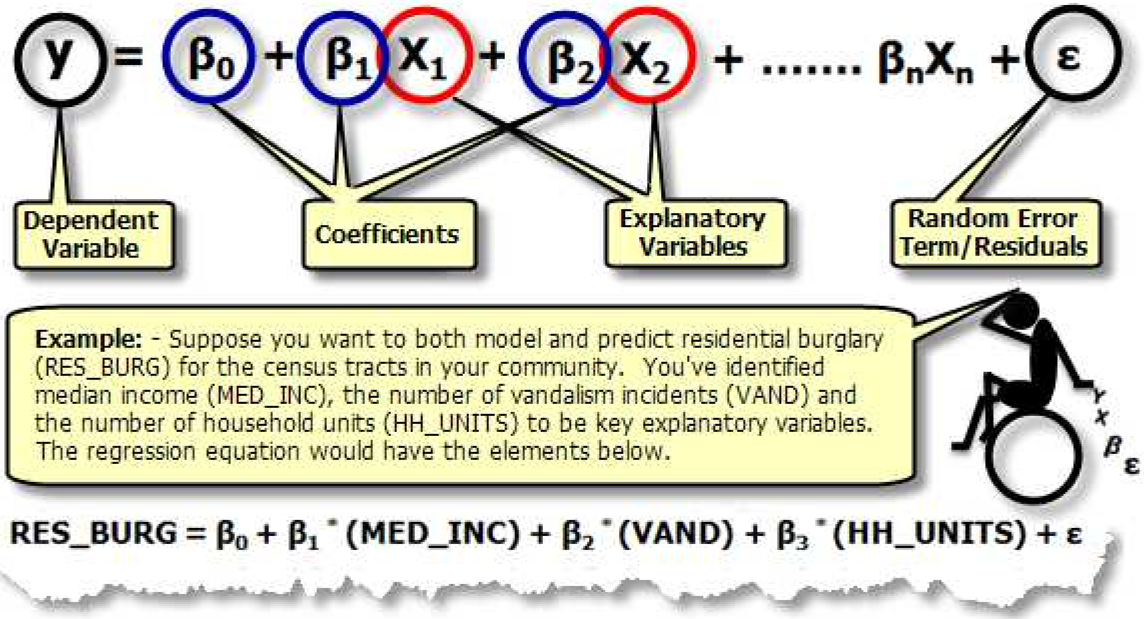


Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. *x* IS CALLED THE *INDEPENDENT* OR *PREDICTOR* VARIABLE, AND *y* IS THE *DEPENDENT* OR *RESPONSE* VARIABLE. THE *REGRESSION* OR *PREDICTION* LINE HAS THE FORM

$$y = a+bx$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ \beta_n X_n + \varepsilon$$



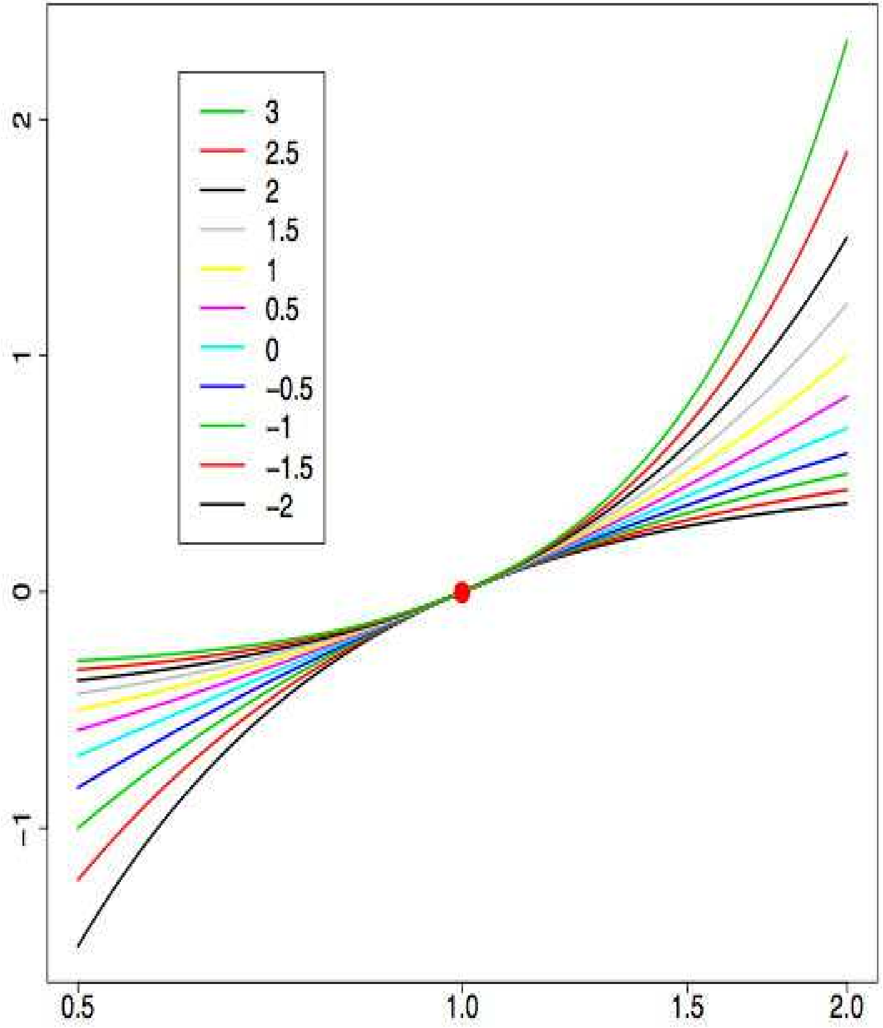
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ \beta_n X_n + \varepsilon$$

Example: - Suppose you want to both model and predict residential burglary (RES_BURG) for the census tracts in your community. You've identified median income (MED_INC), the number of vandalism incidents (VAND) and the number of household units (HH_UNITS) to be key explanatory variables. The regression equation would have the elements below.

$$RES\_BURG = \beta_0 + \beta_1{}^{*}(MED\_INC) + \beta_2{}^{*}(VAND) + \beta_3{}^{*}(HH\_UNITS) + \varepsilon$$

Box-Cox Regression (BC)

- Extension of linear models (lines).
- Now we consider curves with different shapes.
- Uncertainty varies according to predictor magnitude.
- Better fit compared to linear model.
- Problem with change of scale in search volumes.



First experiment: Methane Fueled Cars

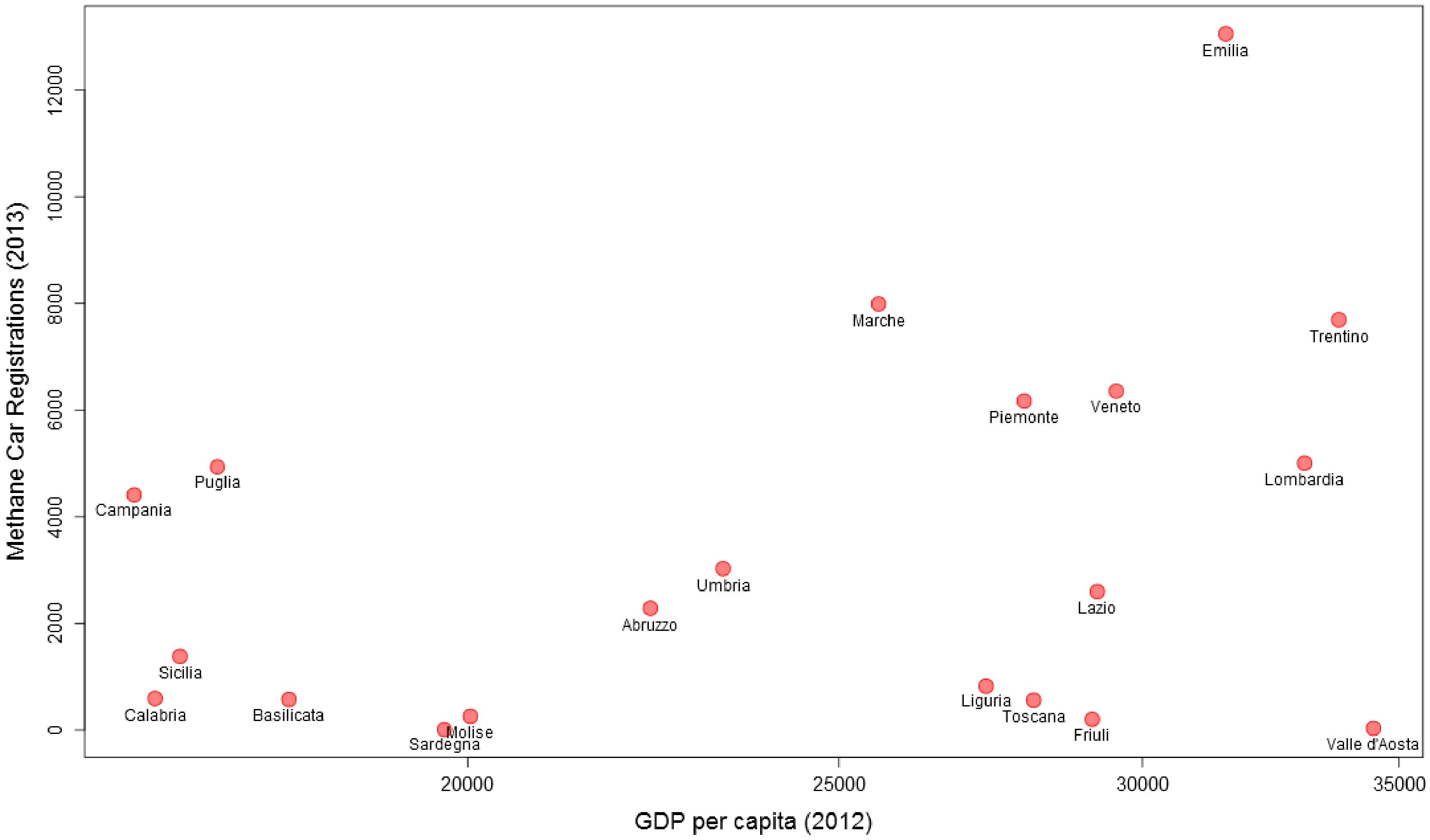
Registrations of Methane Cars (2013)

- Italian car data available on the Automobile Club d'Italia (ACI) website **www.aci.it**
- Split by type of fuel supply and Region.

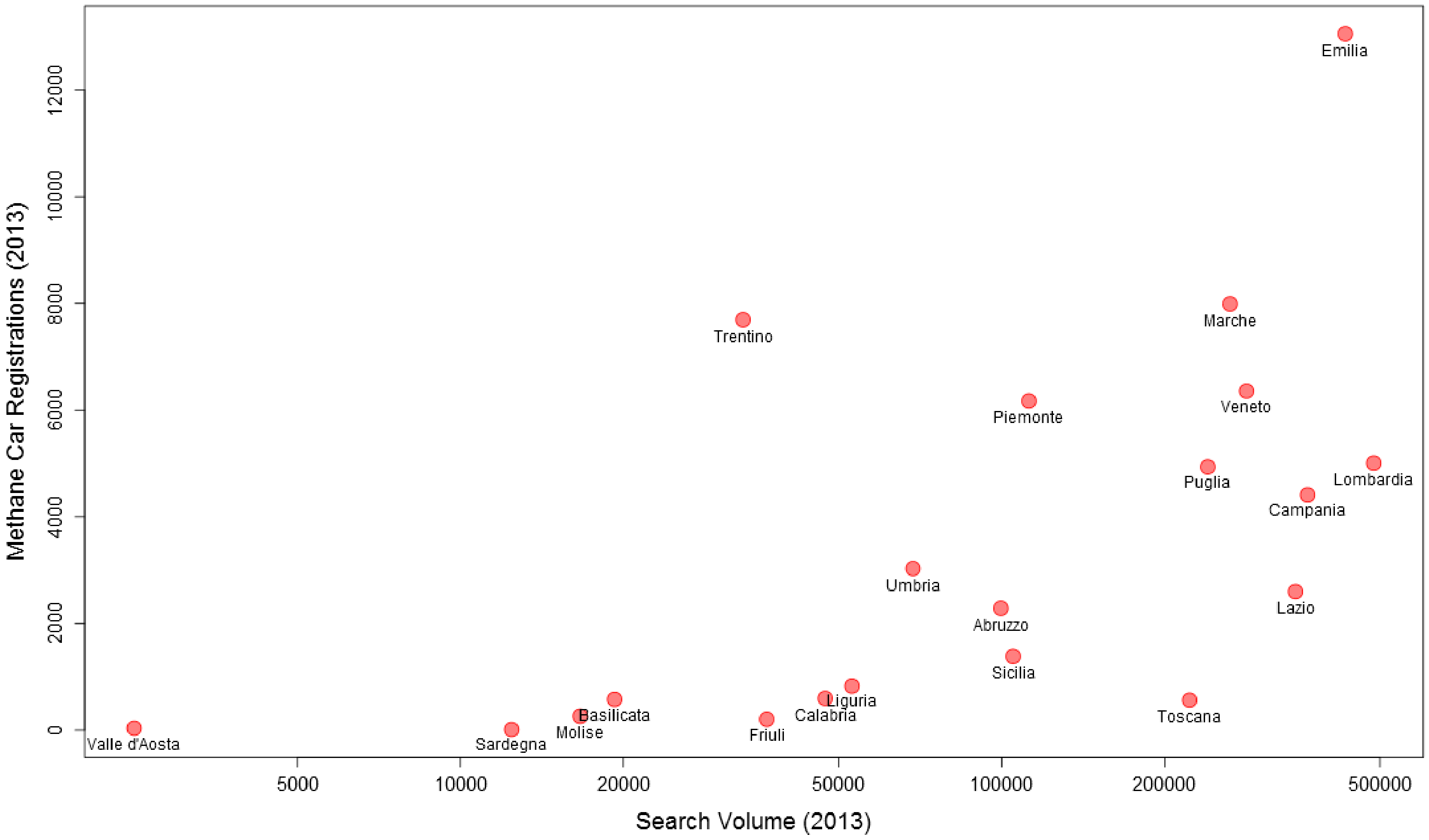
Covariate	Correlation with Registrations
Search volume (Google AdWords)	***0.66***
Population (ISTAT* data 2013)	0.36
Density (ISTAT data 2013)	0.27
GDP per capita (ISTAT data 2012)	0.36

*National Institute of Statistics: **www.istat.it**

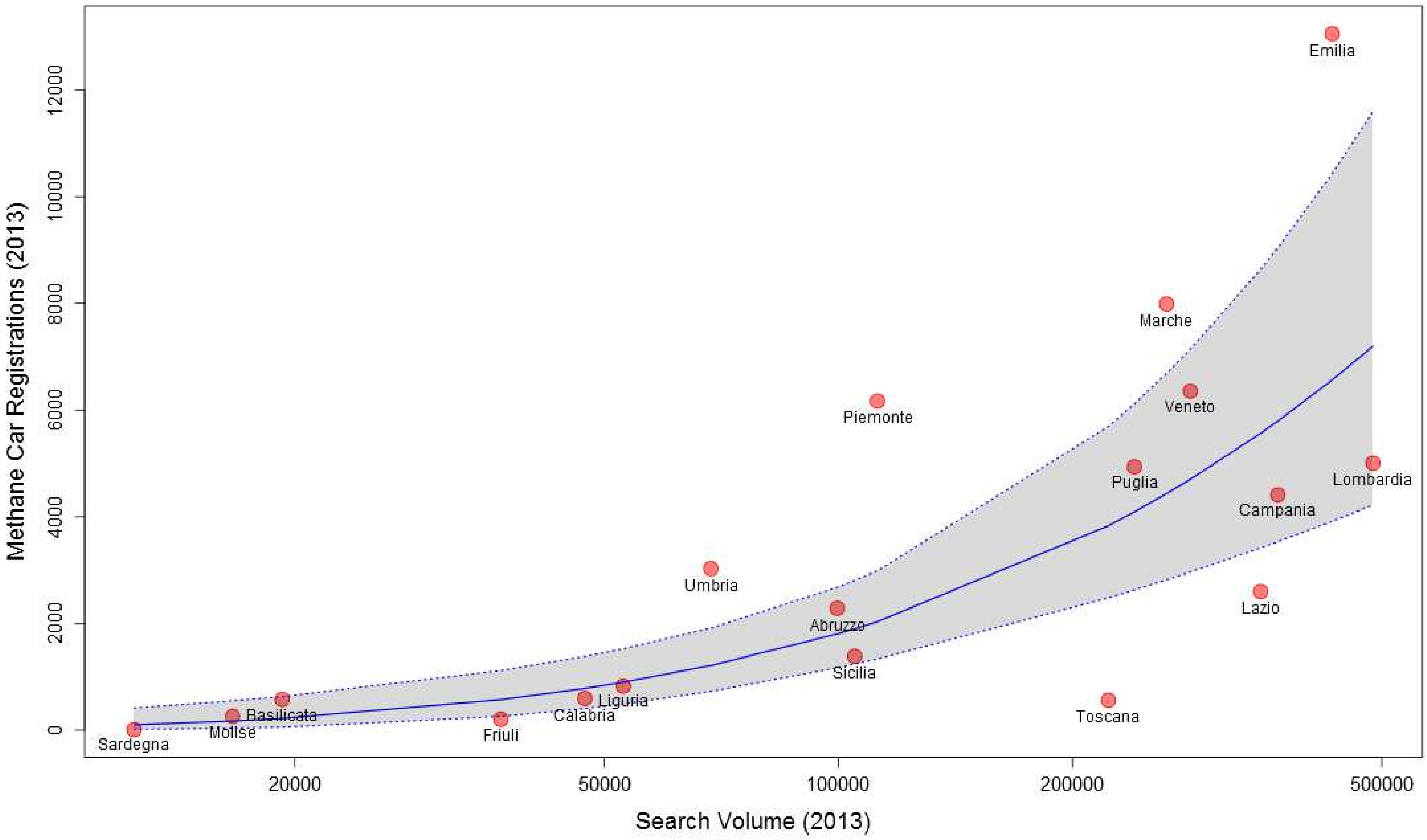
Registrations vs GDP



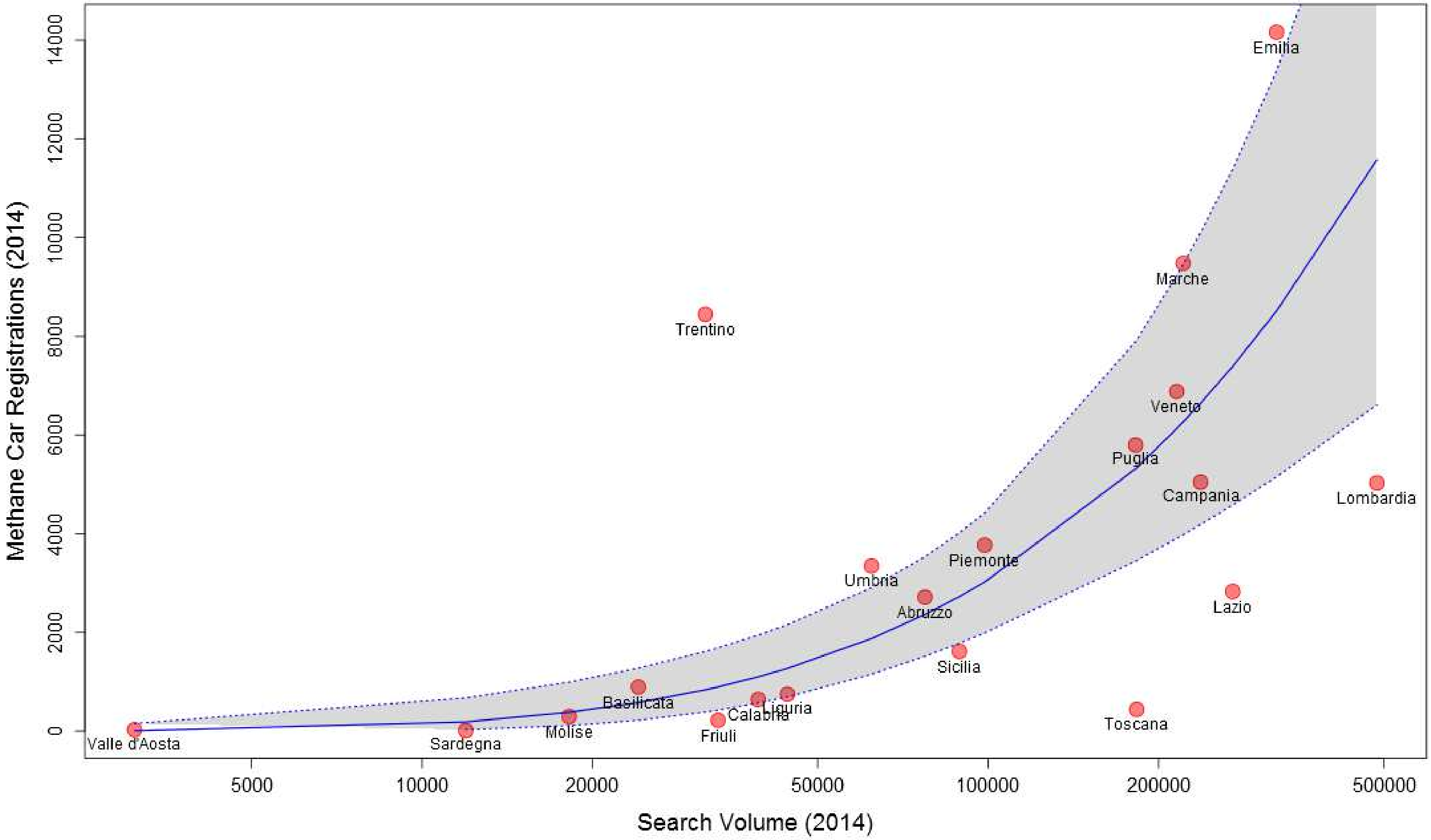
Registrations vs online searches



Methane2013: BC regression, no outliers



Methane2014: fit using Methane2013



Equations (Methane)

$$Regist. = \alpha_{MET}(\log SearchVolume)^{k_{MET}} + \varepsilon$$

- k_{MET} controls the shape of the curve
- $R^2_{MET2013} = 0{,}50$ (Autonomous regions excluded)
- $R^2_{MET2014} = 0{,}48$ (as above)

Second experiment: Photovoltaic Installations

New Photovoltaic Systems (2013)

- Italian installation data available on the Gruppo Servizi Energetici (GSE) website Atlasole **atlasole.gse.it**
- Data on systems subsidized through Conto Energia (CE).
- Split of installation number and power by **Region** and Province.

Covariate	Correlation (N.)	Correlation (Pow.)
Search volume	0,80	0,68
Population	0,84	0,75

Photo2013 (Regions): fit using Methane2013

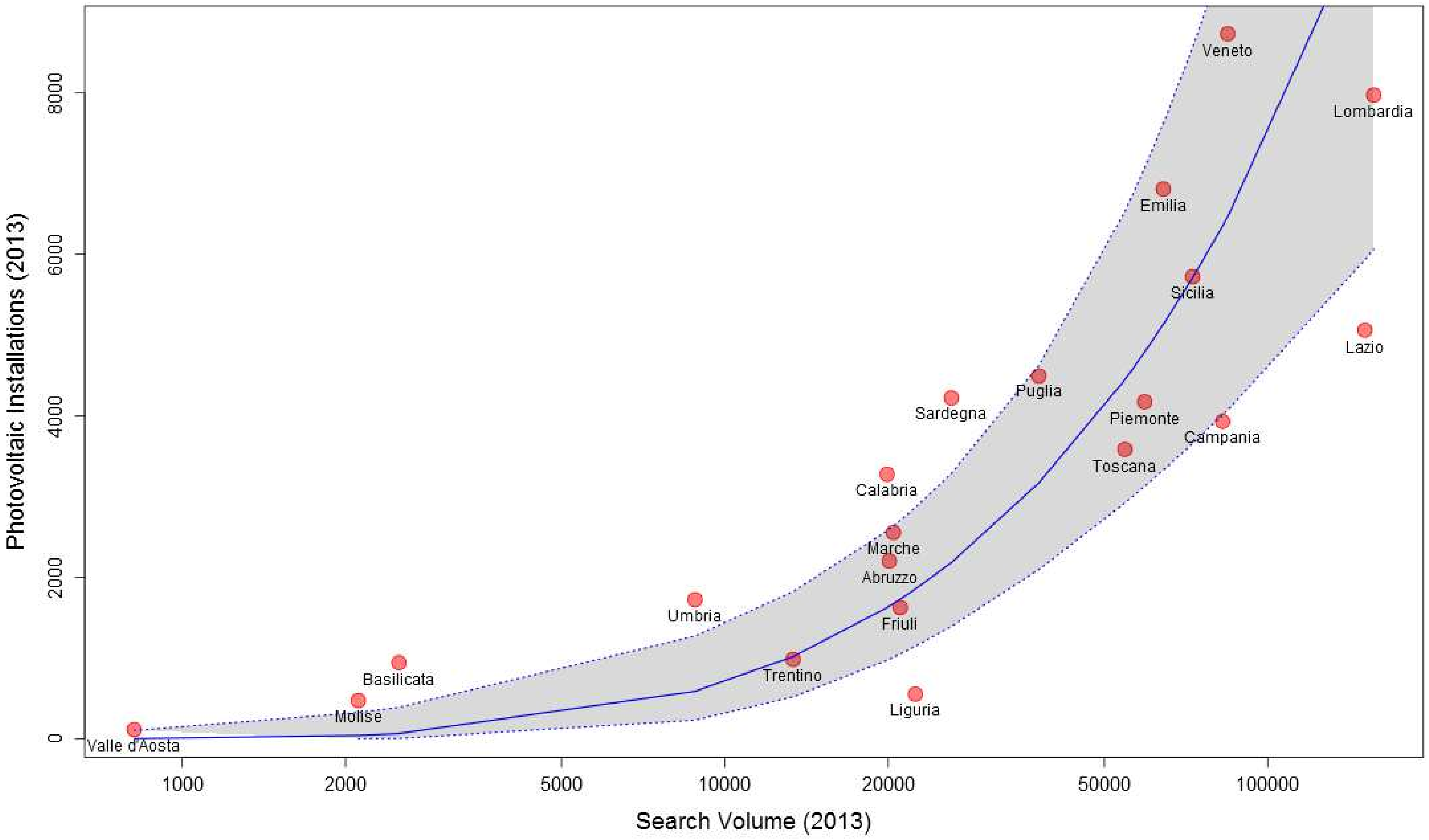
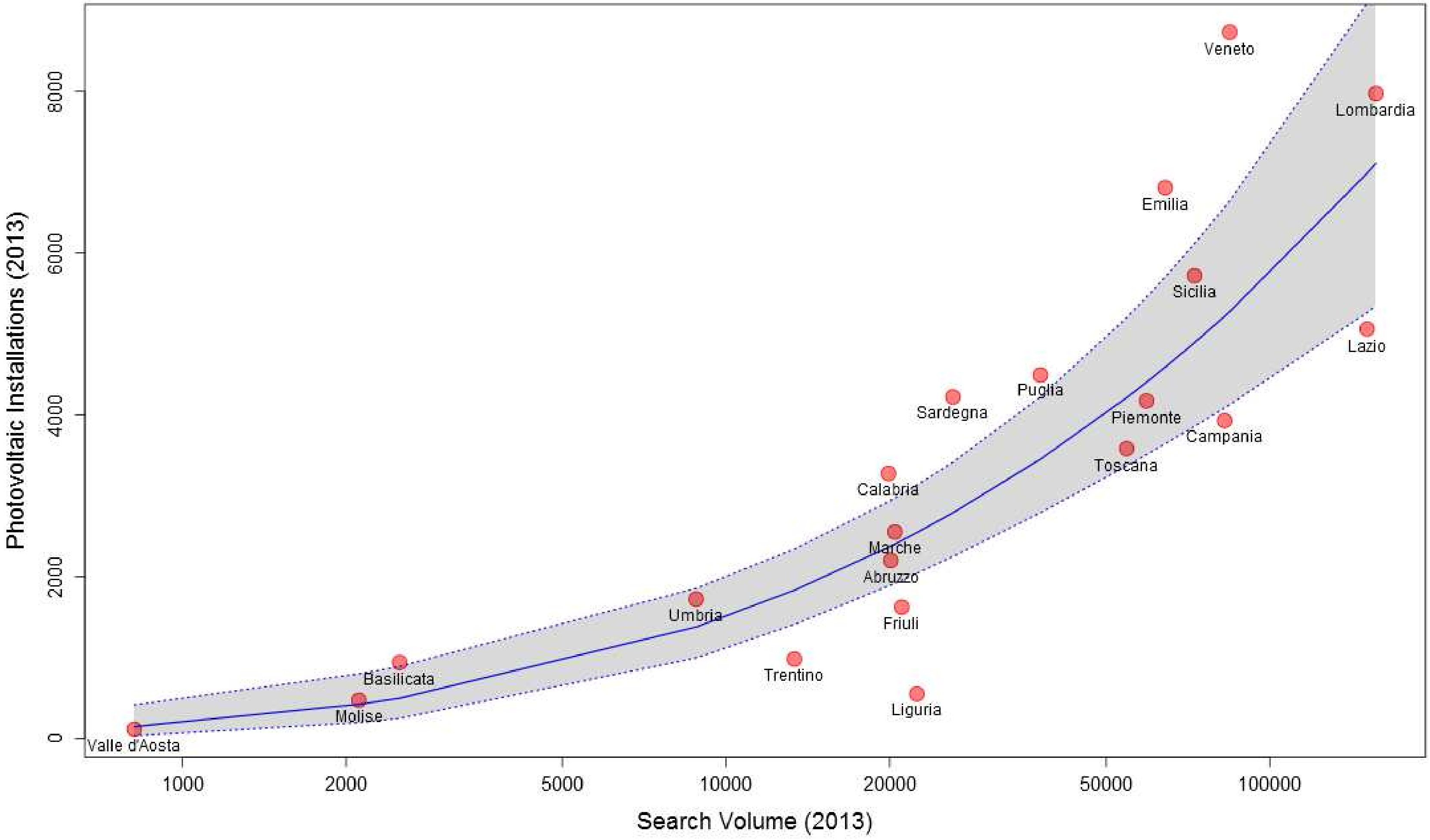


Photo2013 (Regions): BC regression



Photovoltaic Systems (Provinces)

- Model performance drastically decreases in moving from Regions to Provinces.
  - Lower correlation with search volume and population.
  - Why?
- Hypothesis: typology of urban fabric matters.
  - Urban fabric density: number of housing units per building (source: ISTAT data 2011).
- Search volumes, population and urban fabric density provides complementary information on a smaller scale such as Provinces.

Photo2013: from Regions to Provinces

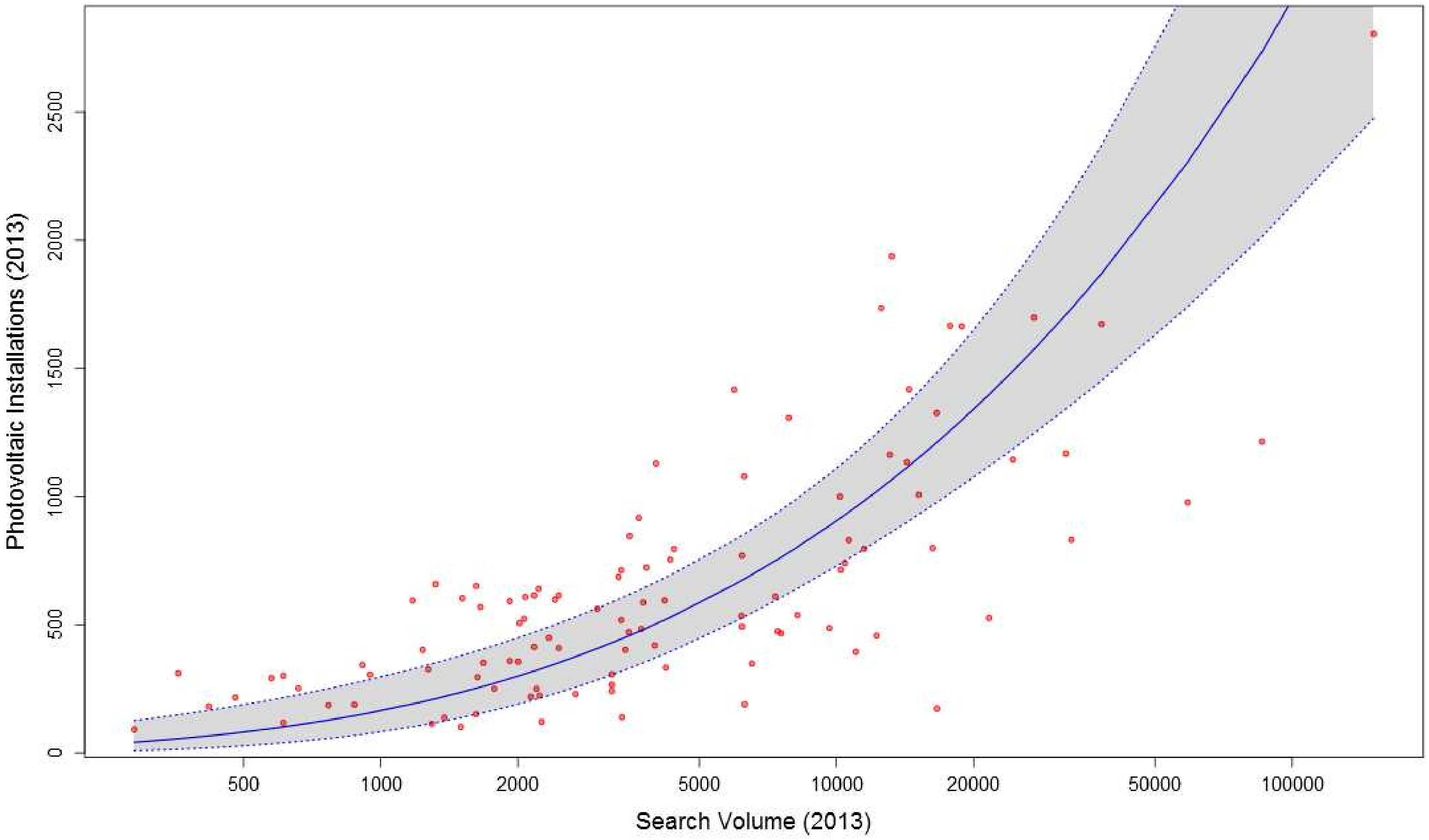


Photo2013: Administrative Centers

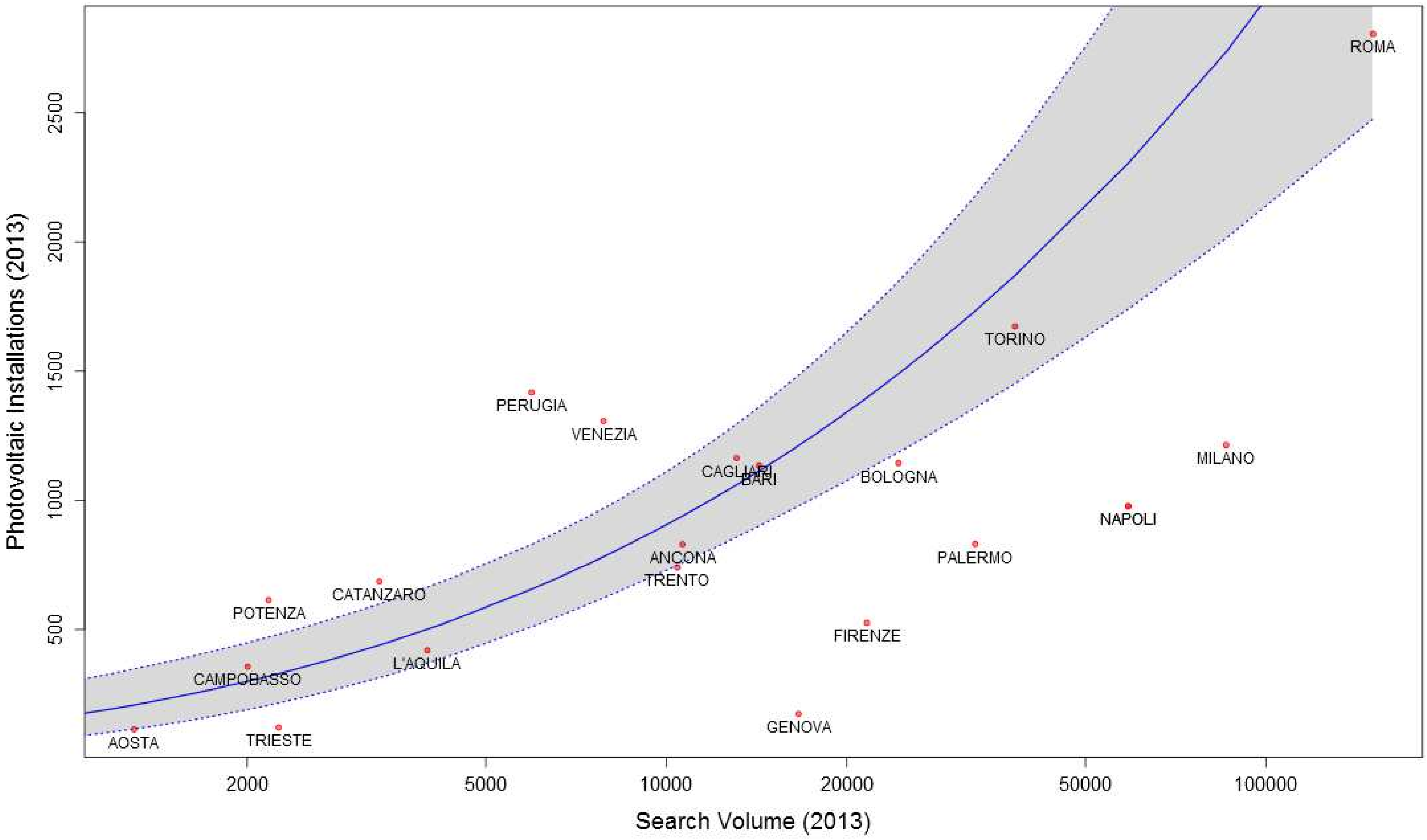


Photo2013: Actual vs Predicted

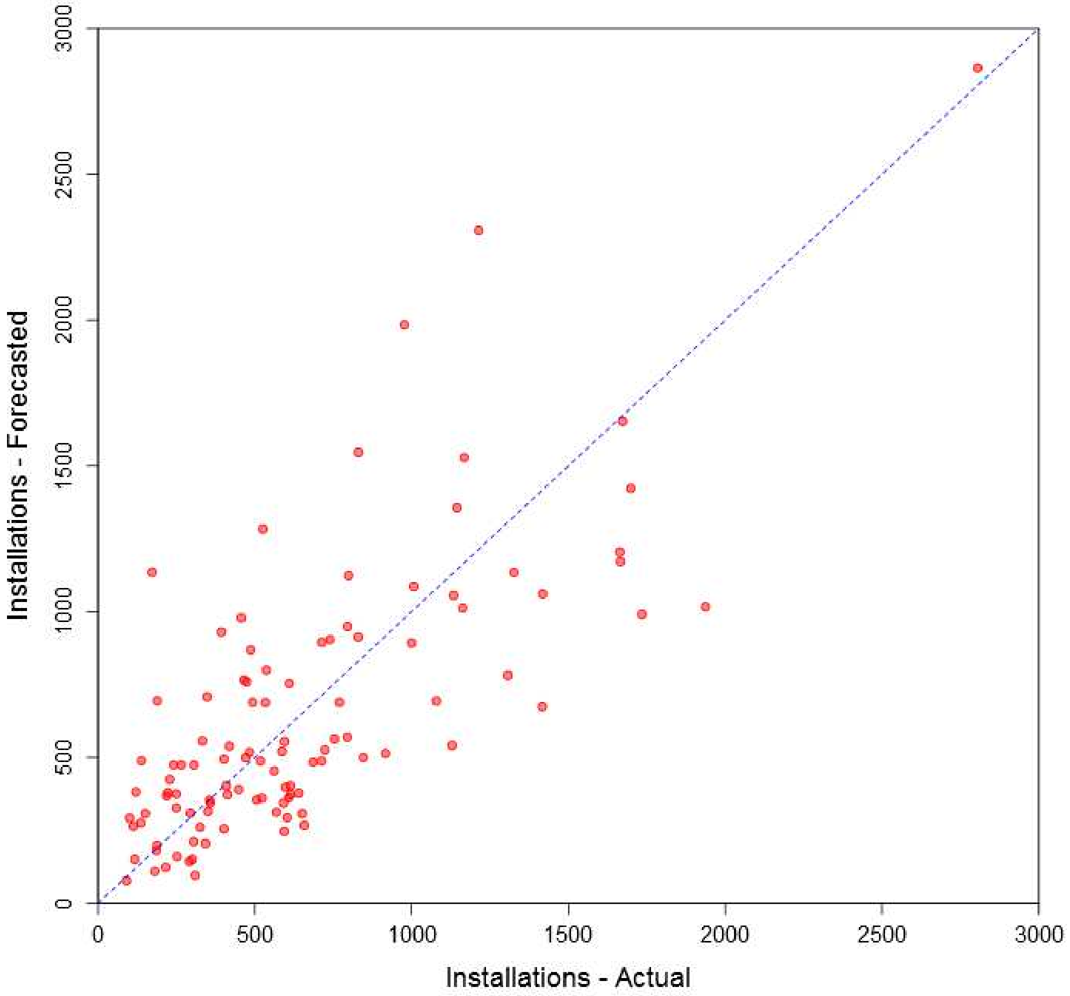
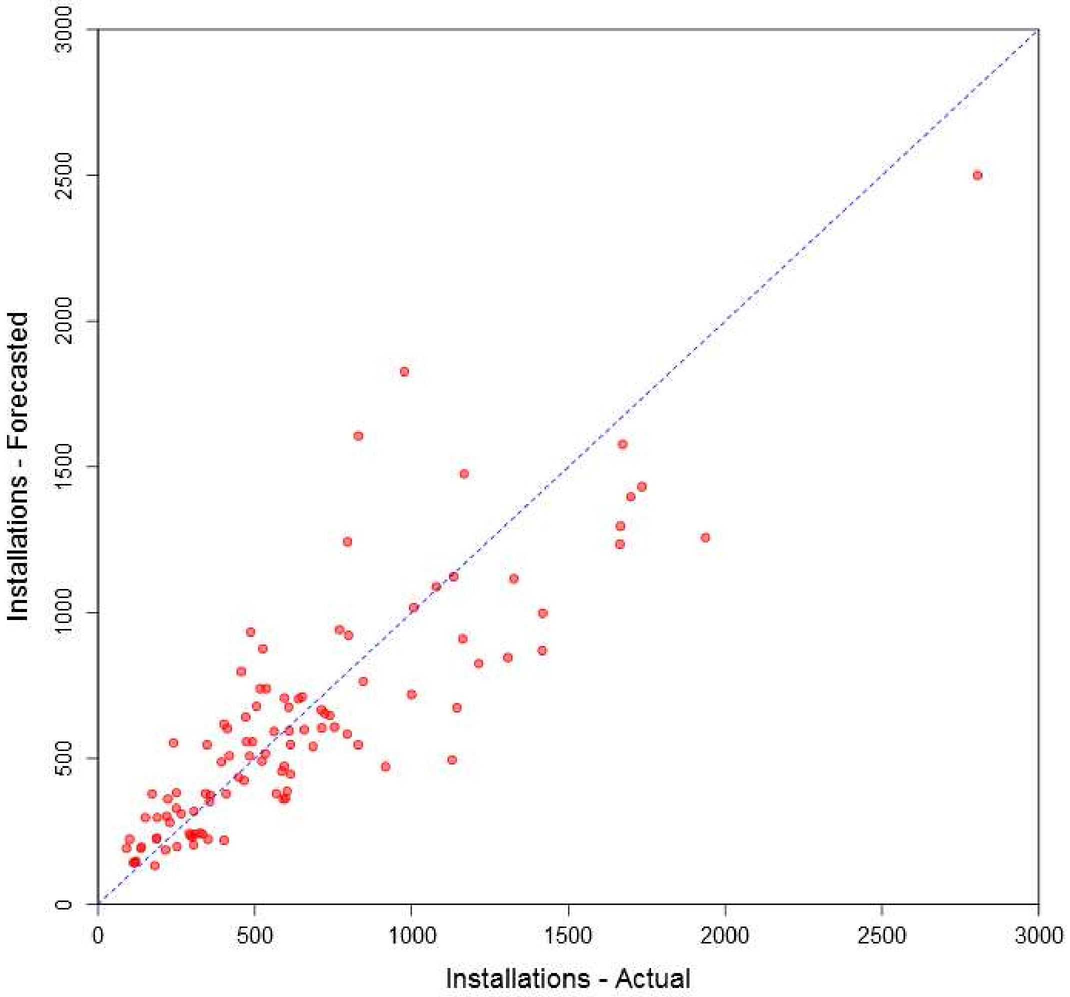


Photo2013: exploiting Urban Fabric



Equations (Photovoltaics)

$$Install.N = \alpha_{PV}(\log SearchVolume)^{k_{PV}} + \varepsilon$$

- $R^2_{REG} = 0{,}72$
- $R^2_{PROV} = 0{,}35$

$$Install.N = \big(\alpha_{PV} \log SearchVolume + \\ +\beta_{PV} \log Population + \gamma_{PV} FabricDens\big)^{k_{PV}} + \varepsilon$$

- $R^2_{PROV} = 0{,}70$

What we have learned so far

- Strong correlations.
- Predictive models.
- Predictive power in time.
- More challenges in change of scale, they can be tackled.

Third experiment: Solar Thermal Installations

New Thermal Systems (2013)

- Reference data:
  - requests for **tax relief** and **average installed surface** by Region (55% relief, *source*: Rapporto Annuale Efficienza Energetica, compiled by ENEA*);
  - Italian total glazed surface (*source*: Solar Thermal Markets in Europe report, compiled by ESTIF**).
- A tentative split of total glazed can be made.
  - Requests are estimated to account for about 35% of total installations.
  - Regional propensity to perform installations without requesting tax relief is not accounted for at this stage (more on this later).

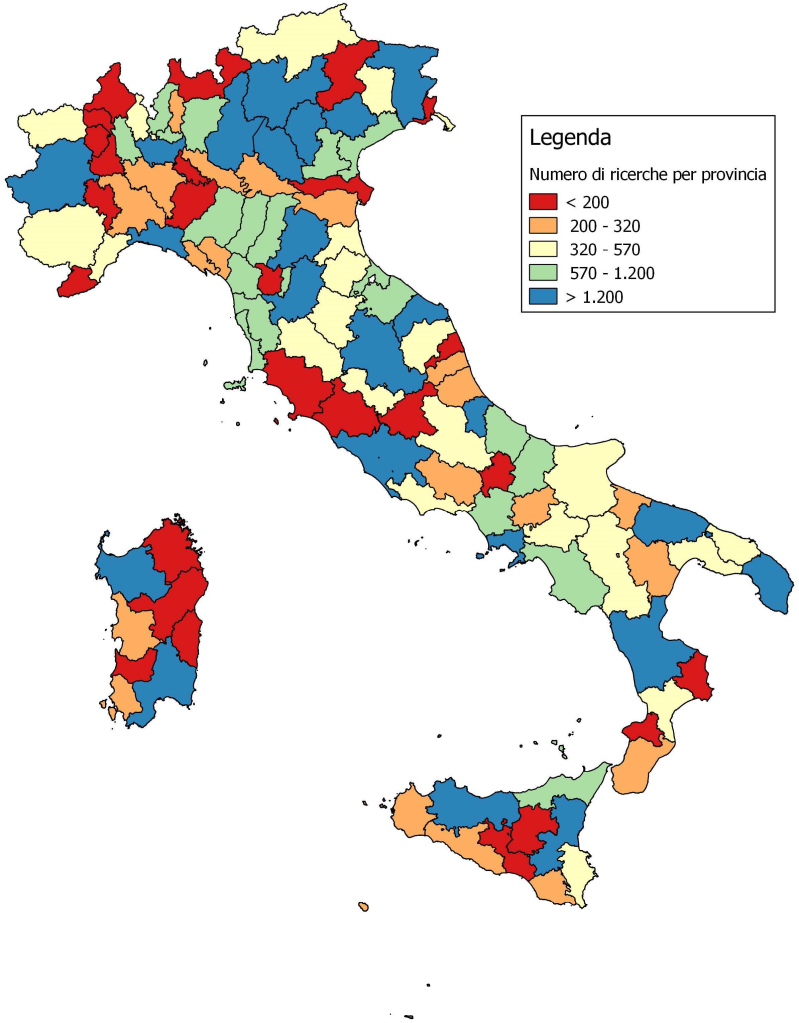
*Italian National Agency for New Technologies, Energy and Sustainable Economic Development, www.enea.it

** European Solar Thermal Industry Federation, www.estif.org

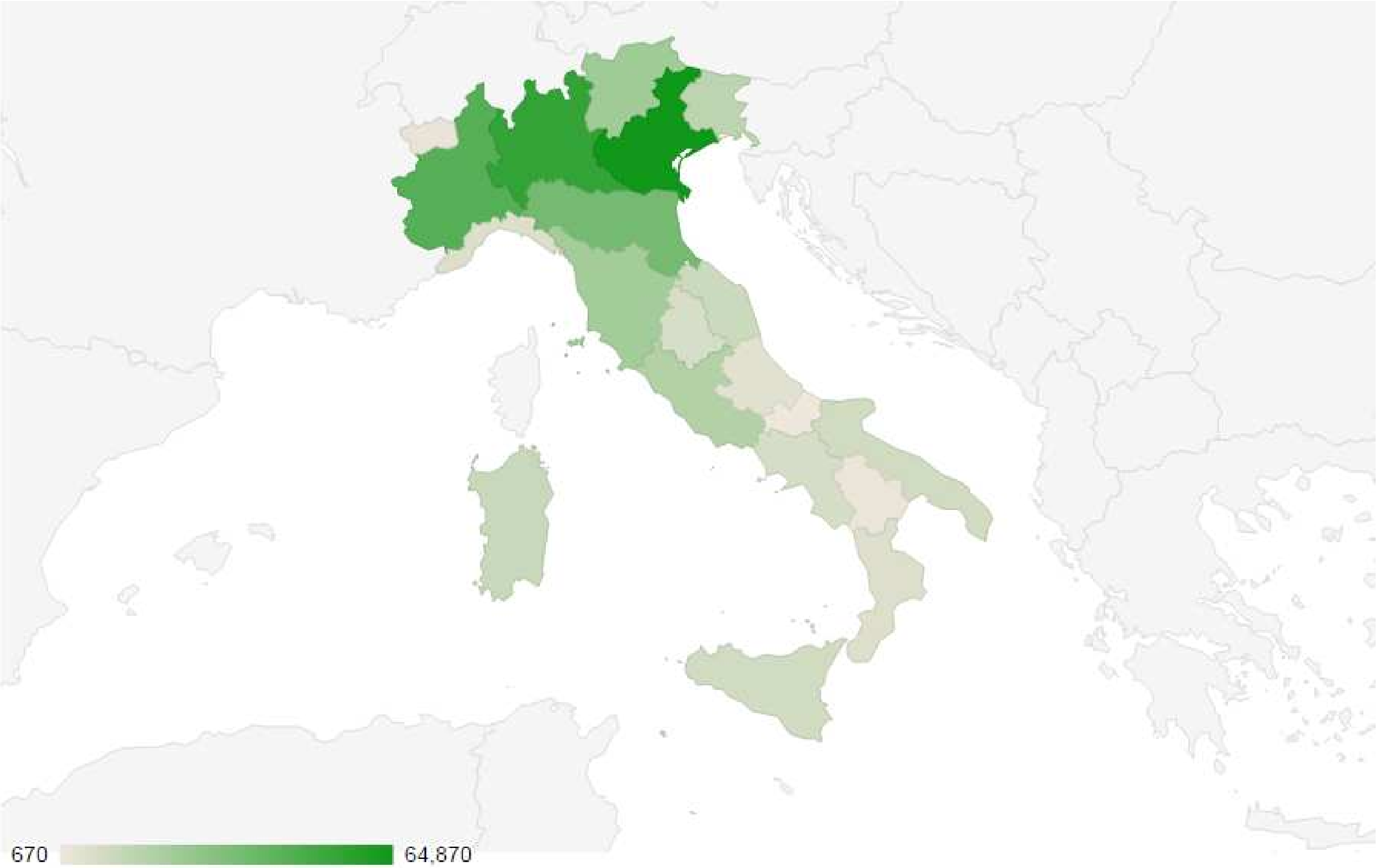
Thermal Installations (Regions)

- Predicted shares of glazed surface using Photo2013 appear very different from estimates made through tax relief requests.
- Relative shares in Northern Italy are very close to what we get from tax relief estimates.
  - Most disagreement is between estimates in Central-Southern Italy.

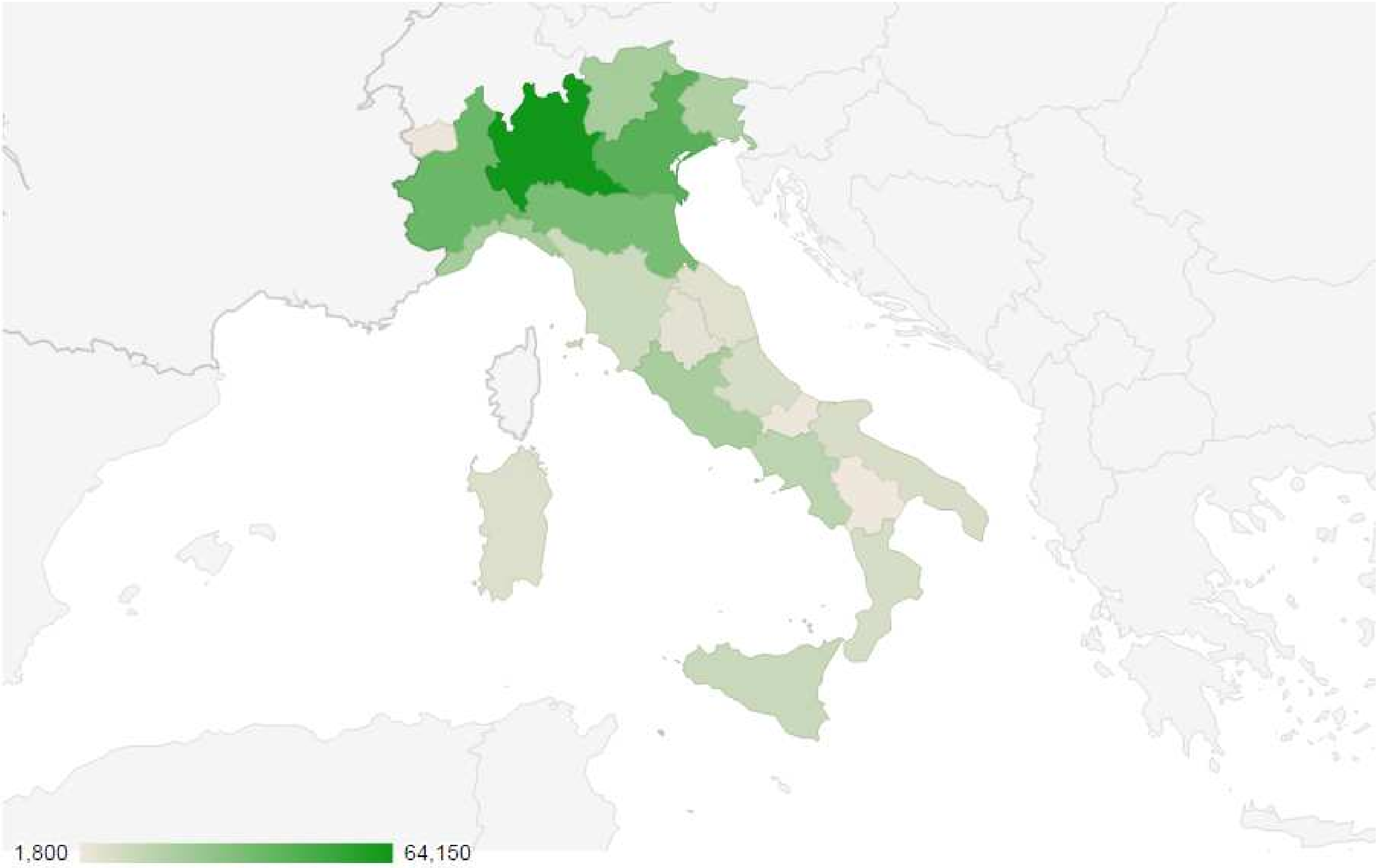
Solar Thermal: avg. monthly searches (2013)



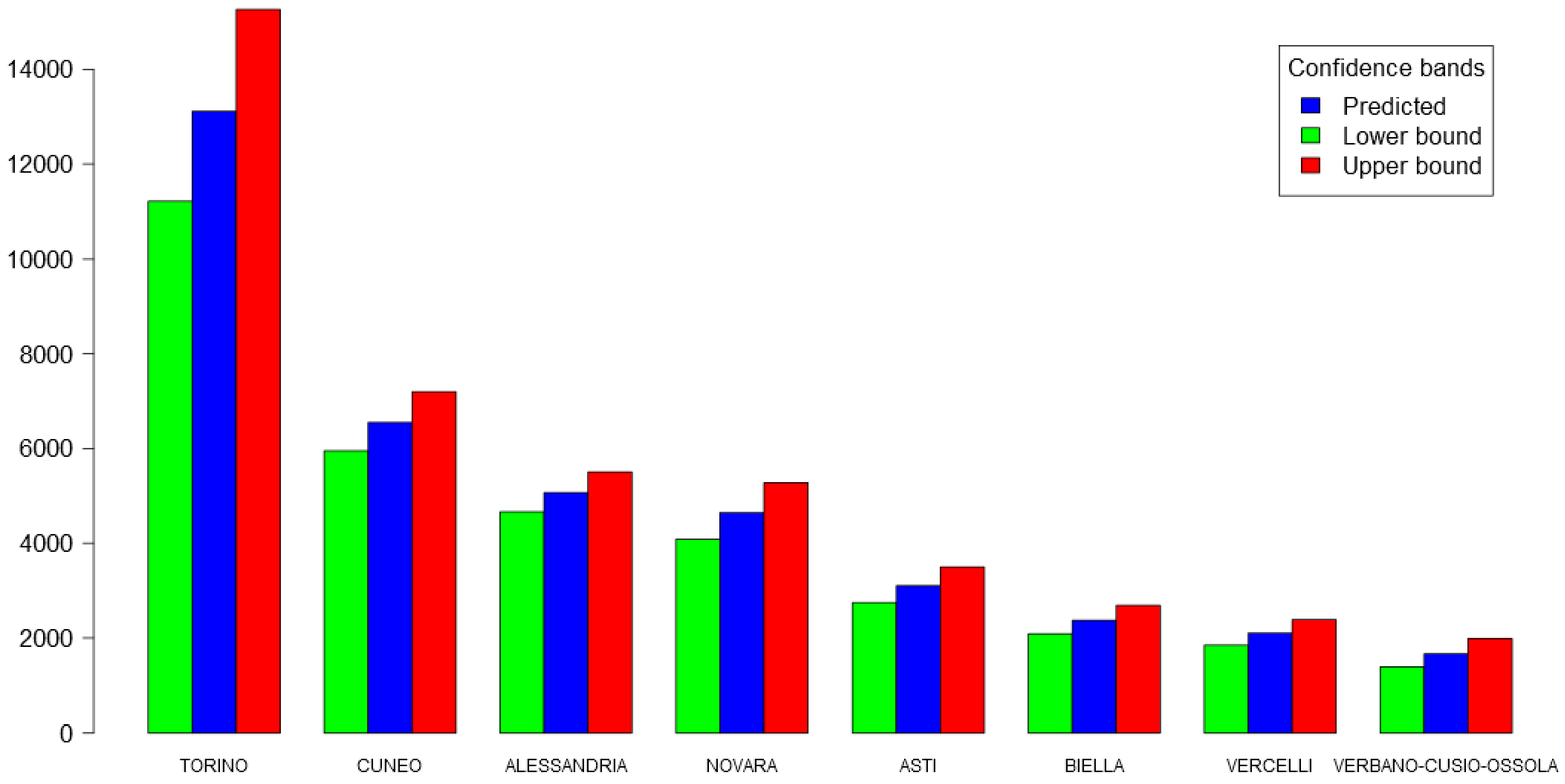
Estimates (2013): National total split according to financial data



Estimates (2013): Separate North/South split according to searches



Piedmont: estimates with confidence



Conclusions

- "Real" attitude towards sustainable energies is reflected in online behavior.
- Online searches provide an additional tool for first estimates, that can corroborate or question existing forecasts.

Torino, 19 novembre 2015

Thank you

Francesco Tarasconi
tarasconi@celi.it
Vittorio Di Tomaso
ditomaso@celi.it

www.celi.it
@celi_nlp