Patnerships for Innovation. Spanish experience

Ricardo Cao

rcao@udc.es, https://dm.udc.es/staff/ricardo_cao/

Research Center for Information and Communication Technologies (CITIC), https://citic.udc.es/

Universidade da Coruña, https://www.udc.es/

Session on Statistical Innovation in the European Statistical System,
INE, Madrid, October 23rd, 2023

Outline of the talk

1. INE research projects call
2. Project DSEIOSF to be carried out at CITIC (UDC)
3. About CITIC
4. Research lines
5. Lessons learned so far
6. Conclussions

Research projects call by INE

Call published in 2021

Budget: 1 M €

Grants for carrying out scientific research projects in cooperation with INE, the Spanish Statistical Office.

Thirteen priority research lines, established by INE, that may be funded:

1. **Automatic coding** with **Machine Learning** techniques.
2. **Natural language** and artificial intelligence applied to **data collection**.
3. **Long-term economic time series:** problems of modeling the **economic crisis** and series **change points**.
4. Use of **conversational systems** in the **dissemination of statistical data**.
5. **Publication and** use of **linked open data**.
6. **Anonymization** of **microdata** and **tables**.
7. Application of estimation in **small areas** for the use of auxiliary information in the **disaggregation** of survey information.
8. Use of **public information on the web** to determine **characteristics of companies** related to innovation and/or the information society.

9. Use of **bank card payment information** to **distribute the expenditure** of non-resident **tourists** and excursionists assigned to the Autonomous Community of the main destination of their visits to Spain among all the Autonomous Communities visited.
10. Study to create a **new panel of companies.**
11. Estimation of **series** within the framework of the indicators of the **2030 Agenda for Sustainable Development**.
12. Use of **benchmarking** to estimate **industry turnover indices** by markets.
13. Calculation of **imputations** in **turnover indices** in the industry through using **machine learning** techniques.

The Spanish Research Agency (AEI) took part in the scientific and technical evaluation.

AEI issued a first report for the assessment committee (formed by INE officers) in order to evaluate:

a) Scientific-technical expected contribution (25%).

b) Ability of the research team and recent contributions related to the project area (25%).

c) Previous results of the research team and assessment of previous activities and projects (20%).

d) Participation of the research team in EU projects or in other international programs or collaborations with international groups (15%).

e) Viability, methodology, research design and work plan (15%).

f) The thematic priorities, lines and sublines in the call.

Projects funded

Final list of financed projects, ranked by the average score along all the lines included in every project

University	**Amount**
Universidad Miguel Hernández de Elche (UMH) Line 7	96.600,00 €
Universidad Politécnica de Madrid (UPM): Line 5	173.650,00 €
Universidad de Granada (UGR): Lines 8, 9, 13	67.045,00 €
Universidad de A Coruña (UDC): Lines 1, 3, 5, 7, 8, 9,13	648.700,42 €
Total	**985.995,42 €**

UNIVERSIDADE DA CORUÑA citic

The project presented by the University of A Coruña (UDC) – CITIC.

Data science and engineering for the improvement of the official statistical function (DSEIOSF)

Project to be carried out in the period May 2023 – May 2026.

Total budget: ~ 649,000 €
Personnel: 72,1%, travels and attending conferences: 10.8%, computer equipment: 2.1%, overheads: 15%

Staff: 23 senior researchers, 3 early-stage researchers, 4 administrative or technical assistants (all these 30 people at CITIC-UDC), 3 senior researchers: 1 in Universidad Politécnica de Madrid, 1 in Université de Genève, 1 in IGE (Galician Statistical Office) and 7 other early stage researchers (predoc) or experience researchers (posdoc) to be hired.

The project addresses 7 of the 13 research lines included in the call by INE: lines 1, 3, 5, 7, 8, 9, 13.

About CITIC

https://www.youtube.com/watch?v=7ocXNdcTMIU

CITIC is a Singular Research Center created in 2008 that promotes progress and excellence in R&D&I applied to ICT. Created by the Universidade da Coruña, CITIC is a meeting point between the University and the ICT sector: R&D departments of companies come together with researchers from the university.

UNIVERSIDADE DA CORUÑA citic IN e

In 2016, CITIC received the distinction of Singular Research Center of Galicia 2016-2019. It was the first and only ICT center in Galicia (NW Spain) recognized with this distinction, co-financed by the Xunta de Galicia and the European Union through the ERDF.

In 2019 CITIC renewed its accreditation, becoming a Research Center of the University System of Galicia 2019-2022. CITIC scientific activity is structured into five research areas, Artificial Intelligence, Data Science and Engineering, High Performance Computing, Intelligent Services and Networks, and Cybersecurity.

Line 1.- **Automatic coding with machine learning techniques**

There are **two types of categorical answers** in a survey:
Closed questions. Answer must be chosen from a predefined set of categories. PRO: Easy to process. CON: Constraining.
Open-ended questions. Free answer. Subsequently, its transcription is manually classified according to predefined categories (survey coding). PRO: Provide more and better information. CON: Very costly and potentially subjective analysis.

Our approach:
Build an **automatic encoder** that, taking (1) the (open-ended) question, (2) the (free) answer, and (3) the predefined set of categories, **computes the most probable categories for the answer**.

We will approach survey coding as a **text classification process**: the answer text is classified according to its corresponding available categories.

Machine learning and **deep learning** state-of-the-art techniques **will be applied**, and their suitability will be analyzed based on
(1) characteristics of the question and answer texts, and
(2) availability (or not) of training sets, i.e, previous surveys already coded.

The **initial use case** will be **economic activity surveys** (National Classification of Economic Activities - CNAE), of two types: aimed **at companies** and aimed **at individuals**.

UNIVERSIDADE DA CORUÑA citic INE

Line 3.- **Economic time series of long duration in time: modeling the economic crisis and series change points**

The aim is to develop a **general methodology** that allows for **seasonal adjustment** in long time series, with **two identified models** and a **transition period** to go from one model to the other. This methodology will be **similar to** that of **TRAMO** (Time series Regression with ARIMA noise, Missing values and Outliers) and **SEATS** (Signal Extraction in ARIMA Time Series), for extracting deterministic effects (including calendar), and filtering to eliminate seasonality, represented **in state space**.

Line 5.- **Publication and use of linked open data** (joint with UPM)

The main aims are:

Publication and **dissemination** of statistical classifications and other structural metadata (for **example**, the National Classification of Economic Activities, **CNAE**), such as **Linked Open Data**. This will facilitate access and understanding of fundamental data for decision-making and analysis.

UNIVERSIDADE DA CORUÑA citic

Development of **an application that adds value** to the open data provided by INE. This application will allow users to **access, visualize and analyze data** in a **more effective and meaningful way**. The application is proposed as an **open data viewer** and will include both tabular, graphic and geographic visualization.

Investigate, design and develop a **query API** that is **flexible, powerful and efficient**. This will allow users and developers to **access data in a personalized and efficient way**.

Establish a sustainable system. The project is committed to creating a long-term sustainable system and process, ensuring that **infrastructure and services continue to be accessible** even after **the research project concludes**. This will help **maintaining** the **availability of data** and **query services** in **the future**.

Line 7.- **Application of estimation in small areas for the use of auxiliary information in the disaggregation of survey information** (joint with UMH)

The main objective is to provide **estimates for variables or domains** of interest where **sample sizes are small** and it does not seem advisable to use direct survey estimators.

Estimated length of this research line is just **2 years**, with one survey to be considered every year. At this moment only the first year problem has been released to us.

The first problem is to provide estimates in the **Labour Force Survey** (EPA) for:

- The **totals** of **employed**, **unemployed** and **inactive** people and the unemployment rate in those municipalities with **more than 20,000 inhabitants**.
- The **total number of NEETs** (Not in Education, Employment or Traning) at **provincial level**
- The **totals** of **employed**, **unemployed** and **inactive** persons and the **unemployment rate by nationality** at national level.

Our initial proposals for the survey to be dealt with in the second year are these three:

The **Household Budget Survey** (EPF)

The **Labor Force Survey** (**new aims** and methods with respect to those of the first year)

The SARS-CoV-2 National Study of Seroprevalence (**ENECOVID**), set up by the National Centre for Epidemiology, with the collaboration of INE for the sampling design and selection.

Line 8.- **Use of public information on the web to determine characteristics of companies related to innovation and/or the information society** (joint with UGR)

The main objective is the analysis and development of machine learning methods that allow us to **collect information available in public websites**, related some of the variables of the questionnaire of the Companies Innovation Survey.

Specific objectives:

Obtaining company website URLs: Automatically detecting a company's website **based on** a few identifying data such as company **name** and **VAT number**.

E-commerce analysis, both from the company website and from marketplaces (amazon, ebay, etc.). Evaluate a list of websites to **conclude** if these companies **sell their products through their website**, via some kind of online store, and/or **through marketplaces**.

Chat analysis. Examine a list of web sites to determine if they have a chat where one gets a **response from a human** agent or an **automated chatbot**.

Innovation Analysis. Detect whether **a company is innovative** based on the presence of **certain keywords on its web pages**.

Obtaining business turnover indices. Performing search tasks in publications on the company’s web pages or using public records to collect such information.

Planned tasks:

Analysis of the information available in INE about the companies under study and the aggregate/disaggregated indicators that should be obtained. Extract the most relevant parameters (name, identification code, acronym, web, address, sector, etc.) that will later be used as input for the systems to be implemented in the project.

Analysis and implementation of web scraping models for the extraction of the specified metadata, implementing protocols that reduce the possibilities of blocking (speed optimization and crawling periods, batch processes, multi-threaded systems, etc.). Study the use of machine learning algorithms to improve the classification of text data on the web site and in the recognition of patterns within the HTML structure.

Evaluation of the results obtained by the models developed through the use of a validation of the models on the studies available in the Survey.

UNIVERSIDADE DA CORUÑA citic INE

Line 9.- **Use of bank card payment information to distribute the expenditure of non-resident tourists and excursionists assigned to the Autonomous Community of the main destination of their visits to Spain among all the visited Autonomous Communities** (with UGR)

A methodology will be developed to **distribute the expenditure of tourists and excursionists** at the destination estimated with **Tourist Spending Survey** (EGATUR) and **assigned to** the Spanish Autonomous Community of **main destination**, among all the

Autonomous Communities visited on the trip or excursion.

To do this, **additional information** that other auxiliary sources can provide will be used. More specifically, information to be used is related to **in-person payments** through point of sale terminals (**POST**) and cash withdrawals at automated teller machines (**ATM**) made in Spain by **card users issued by foreign banks**. This auxiliary information is provided in an **aggregated way** and consists of **payments** performed **each month** and in **each province** by tourists, **grouped by nationalities**.

Line 13.- **Calculation of imputations in turnover indices in the industry using of machine learning techniques** (joint with UGR)

The **turnover index** (ICN, índice de cifras de negocio) is **calculated monthly** for the approximately **12,000 industrial establishments** distributed throughout the country, based on certain productivity, location and activity variables. These data are collected every (approximately) 50 days. The work to be carried out by the UDC team can be summarized in two tasks:

Use of **machine learning** techniques for the **imputation of data involved in the calculation of the total ICN** and disaggregated by markets, in order to be able to **bring forward the publication date** of the official statistics without having to wait for real data.

Manual and automatic construction of the regressors involved by machine learning, analyzing whether **useful combinations (regressors)** different from those derived by the experts are constructed **in the layers** of the machine learning models.

The **UGR team** will be in charge of **analyzing a possible reduction in the sample**, a **change in the sample size** or a **change in the collection periodicity** by reconstructing the total sample with the use of machine learning techniques as well as performing **back-casting**.

Lessons learned so far

Too lengthy evaluation-and-selection process for the projects to be funded

Rethink about the selection process to adopt a by-research-line approach (rather than by project)

Start contact and preliminary work once the provisional list of funded projects is ready

Speed up the hiring process at universities

Try to avoid mismatches between the methodology proposed by the research teams and the one that INE “had in mind”

Use a more interactive procedure for the call, for instance a public procurement of innovation (PPI) approach

Try a *looking for a head that fits to this hat* approach for a few lines and not only a *looking for a hat for this head* one as a unique method for all the research lines

Conclusions

A very interesting experience that places official statistics needs and data science research at the forefront of cooperation

A useful tool to solve important practical problems in official statistics

A first step towards a promising more intensive collaboration between official statistics offices and universities (or research centres)

Thank you for your attention!

You can contact me at … rcao@udc.es

https://dm.udc.es/staff/ricardo_cao/
https://citic.udc.es/

Session on Statistical Innovation in the European Statistical System,
INE, Madrid, October 23rd, 2023