The French SSP Lab: bringing it to life

Dr. Elise COUDIN, Dr. Mathilde POULHES
INSEE, FRANCE

Abstract. INSEE created a new unit dedicated to innovation and R&D in terms of new data sources and new statistical methods for the production of official statistics. Part of the Directorate of Methodology, Statistical Coordination and International Relations, the SSP Lab (Official Statistics Service Lab) is a resource and animation centre for applied research, experimental development and new ways of working within the French official statistics service (*Service de Statistique Publique;* SSP). Sponsored by business units in charge of statistical production, the SSP Lab conducts experimental projects in partnership with these units. It also leads networks within the French official statistics service, in collaboration with external partners, including European peers and academics.

1. What is the background of the SSP Lab creation?

1.1 General context

The extraordinary development of new data sources, methods and processing capabilities of the digital revolution requires official statistics to produce ever more detailed information and to improve timeliness. INSEE, like other NSIs, therefore pays close attention to developing and federating innovation within the national official statistics service. Official statistics must absolutely maintain a high level of excellence in order to produce objective, relevant and meaningful statistical inputs in the public debate within a competitive context of new emerging data providers and actors. This level of excellence requires rare skills that must be mobilised and developed throughout the agents' careers, in a general context of reduction of public sector resources.

In this context, INSEE set up in May 2018 an innovation unit called the SSP Lab (Official Statistics Service Lab, SSP stands for *Service de Statistique Publique*, the official statistics service, i.e. INSEE and the Ministerial Statistical Services), which aims to monitor and disseminate innovative statistical methods and explore new data through experimental projects carried out in partnership with business units of the official statistics service.

The creation of this unit corresponds with the ESS vision 2020, which aims to explore new opportunities of the digital transformation and build organisations capable of working and collaborating with agility within the official statistics service and with European peers (e.g. through the Big Data task force and related ESSnets), but also with data producers and academics. It is inspired by the experience of other NSIs in the Netherlands, Italy, Canada and the U.-K. amongst others that have modernised their organisations to respond to these challenges.

1.2 Context in France

The SSP Lab was created following experiments using Big Data sources carried out in several of the INSEE and Ministerial Statistical Services units, e.g. the project incorporating mass retail scan data for producing the CPI (INSEE, started in 2011), the web scraping of job ads for estimating job vacancies (Ministry of Labour), and the use of administrative health data for statistical purposes (Ministry of Health). There has been also a positive regulatory context for exploring new sources of data since the adoption of the so called "Law for a Digital Republic" of 7 October 2016. At INSEE request, the latter makes mandatory the transmission of information from internal databases for companies concerned by a statistical survey.[1]

The SSP Lab was also created following the restructuring strategy of INSEE. INSEE set up in 2012 the Directorate of Methodology, Statistical Coordination and International relations (DMCSI) with the idea of pooling rare and strategic resources for possible synergies. Within the Department of Statistical Methods (within the DMCSI), the Division of Applied Econometrics and Evaluation (DMAEE) explored and disseminated innovative statistical and econometric methods within the official statistics service by providing support and advice to statisticians in charge of production. This division gradually integrated a role of coordination and animation of the work on Big Data within the official statistics service. It included two full-time data scientists in 2016. It joined the network of European peers through the Big Data task force and the ESSnet Big Data. It conducted several experimental projects on new sources, e.g. evaluating the interest of Internet sources for nowcasting economic indicators (Combes, Bortoli, Renault, 2015 and

1 It also cancels data transmission royalties between public administrations for statistical purposes

Combes, Bortoli, 2017), and started a collaboration with the Orange SenSE laboratory on mobile phone data. In response to requests from business units or Ministerial Statistical Services, investments were also conducted to acquire skills and experience on textual analysis and machine learning methods, and to disseminate practical instructions. These investments provided the opportunity to launch reflections on working modes, allowing different profiles to work effectively together. In collaboration with the IT department, a team of two persons from the division and one from the IT department won the second prize of the Big Data Hackathon organised by Eurostat during the New Techniques and Technologies for Statistics conference in March 2017.

Datascience innovation for statistical production and studies is a crucial activity for maintaining high-quality, meaningful and relevant statistics, and is pledge of productivity gains. The visibility of its innovative works is also essential for the NSI to reinforce public confidence. In 2016, INSEE's medium-term strategy (INSEE 2025) recommended the creation of a unit dedicated to R&D for promoting reactive lab testing and experimentation, ensuring a technological watch, as well as leading and animating a network on related topics. Relying on this, a launch committee worked on the organisation and the objectives of such an entity from November 2016 to July 2017, and proposed to set up the SSP Lab unit within the DMCSI, capitalising on the division DMAEE resources and experience. The aim of the SSP lab would not be to concentrate all innovation within the official statistics service, but to be an expert resource and partner catalysing innovation. It would produce in partnership with business units innovative prototypes using new sources, new methods, new tools, even new angles of study, upstream of production projects. It would join innovation networks, within the official statistics service, but also with academics and European peers. All statisticians of the official statistics service are indeed trained at the master-level engineering *Grandes Ecoles* ENSAE and ENSAI, which are specialised in statistics, computational IT for statistics and economics, and have therefore already integrated datascience and technologies for Big Data in their curriculums. As such, a "branched" vision within the whole official statistics service of the datascience innovation can be adopted. The SSP Lab programme relies mainly on the business unit proposals, functions with reactivity (mainly through experimental projects that assume the risk of failure) and one of the important roles of the SSP Lab is the dissemination of a culture of innovation and knowledge, exchange of

good practices, sharing feedback.

The French Decree of 10 April 2018 modified the INSEE organisation to set up this new unit. Closely related, the new Directorate of Information Technologies. This Directorate will include a unit dedicated to IT innovation called the Enterprise Architecture, Security and Innovation (EASI) that will foster technical innovation in the computing field (for example, by setting up distributed computing platforms and containerisation). The SSP Lab and the EASI will work in close collaboration.

2. How does the SSP Lab function in practice?

2.1 Composition, governance and resources

The SSP Lab team is made up of eight full-time datascientists, econometricians and IT specialists, including its chief and deputy. The idea is to group complementary profiles in terms of skills and experience (seniors / juniors).

The governance of the SSP Lab is ensured by the members of the executive committee of INSEE and the chiefs of the Ministerial Statistical Services, who review once a year the performance of the Lab experiments and provide global directions for future programmes. These directions are voluntarily flexible enough to ensure reactivity and innovation on unexpected topics. Consequently, the investment decisions are made continuously.

The SSP Lab works in close collaboration with the IT innovation unit, notably through the use of the new IT platform that allows distributed computing and other treatments adapted to the needs of Big Data manipulation. The current version of this platform is unfortunately still incompatible with the security requirements of individual data. Consequently, only anonymised samples are used in the experiments.

2.2 Activities

The SSP Lab's activities involve experimental projects, networking activities and dissemination.

a. Experimentation

The Lab conducts applied research and experimental developments involving statistical or datascience innovation in partnership with services in charge of production. The experimental

projects are defined as follows. The subject is proposed and sponsored by a business unit. The project should last around six months. More generally, the idea is to define several stages that entail specific deliverables. The approach is exploratory, on a "small" scale, and similar to the lab testing approach. Consequently, the deliverables may be of different types: proofs of concept (POC), reports, experimental prototypes, etc., but cannot be directly integrated into a production process. Ultimately, they may not even be implemented. Experimental outcomes never being certain.

Several conditions must be fulfilled to launch an experimental project involving the SSP Lab and a business unit. As already mentioned, the needs are expressed by the business unit, should be in keeping with the orientations previously mentioned, and should concern (at least) one of the Lab's fields of action (new data sources, new statistical methods or new statistics). In addition, the SSP Lab should have the appropriate skills or should be able to acquire them rapidly.

If the previous conditions are fulfilled, then a mixed SSP Lab/Business Unit/IT team with the required skills (datascience, IT, etc.) is set up. Flexibility and real commitment to the project are required from the different participants. Several models of engagement, e.g. one day-per-week contractual commitments, customised projects and team descriptions are possible.
Flexible and ‘agile’ ways of working are promoted. The mixed team is expected to work in steps and cycles, and to deliver regular outputs. The requirement is to broadly disseminate the results (final and intermediary) and to share experience through different media (intranet, newsletters, etc.).

b. Networking activities

For some experimental projects, the SSP Lab contributes to and benefits from European expertise. In particular, the SSP Lab takes part of the European exploration of the potential of Big Data to integrate it into the official statistics production (Big Data task force and related ESSnets). It participated in the ESSnet Big Data I Mobile Data Working Group, which aimed to clarify the possibilities of accessing mobile phone data, lay the foundations for a methodology for their treatment and estimate population present within a given place and time and related indicators. The SSP Lab will continue to participate in the investments on mobile phone data in

ESSnet BD II and will coordinate the French participation in other work packages, including those on satellite data and smart statistics.

For other experimental projects, the SSP Lab may also form academic partnerships in order to benefit from external expertise. For instance, it currently collaborates with the Institute of Public Policies (IPP) of the Paris School of Economics (PSE) to explore the modelling of professional careers for microsimulation purposes by using machine learning methods. Partnerships with private actors for experimentation on private data are also in the scope of potential activities.

c. Dissemination

The SSP Lab ensures the role of monitoring and dissemination of innovative statistical methods through training on datascience methods (e.g., machine learning, textmining and coding languages such as Python) for the statisticians of the official statistics service and the provision of technical documents. The SSP Lab animates networks on innovative topics within the SSP (dissemination of a Big Data newsletter and Big Data seminars). The SSP Lab and the IT innovation unit also work together to promote 'agile' ways of working, in particular via collaborative workshops and hackathons open to the members of official statistics service and close institutions.

3. Future plans

The short-term plans of the SSP Lab are first to conduct the new experimental projects, increase its visibility within the official statistical service and increase the dissemination of its outputs on appropriate medias (blog, intranet, experimental page on the Internet). A second important issue is to develop appropriate contractual frameworks to host in the SSP Lab external researchers, postdoctoral fellows and PhD students.

4. Some examples of experimental projects

The section details three examples of ongoing experiments. The entire list of activities planned for 2018 is available in the appendix.

4.1 Employer identification in census survey

Sponsor: INSEE Social Studies Directorate (Census unit);
Team: SSP Lab (3 members), Census unit (2 members), IT (4 members), other units (2 members);
Schedule: January 2018 (hackathon) and then from June to December 2018 (experimentation);
Expected deliverables: training (hackathon), an experimental prototype and an experimentation report.

a. What opportunities do Big Data techniques offer?

Currently, respondents of the Census report the name of their employer, the activity of the legal unit and the address of their workplace. These response boxes are filled out in a non-standardised way, and frequently result into incorrect answers (spelling mistakes, imprecision and confusion between fields). In order to obtain a relevant industry code for each job, an automatic coding of employers is currently processed, but it is successful for only 45% of respondents. The remaining 65% are manually coded, requiring the work of around 70 INSEE agents for five months each year. Big Data techniques seem to offer great opportunities to improve this process.

b. Organisation of a Hackathon

The SSP Lab in collaboration with the IT department organised the first Hackathon of INSEE on this subject on 18 and 19 January 2018. It gathered more than 60 persons from the whole SSP (INSEE and Ministerial Statistical Services) and its partners (the Health Insurance Institute-Cnam, the Central Bank, the Employment Agency-Pôle Emploi, etc.). Two days of training were organised before the Hackathon to present the subject and the approach. Different speakers presented the data involved (the Census and the business register, called SIRENE) and some techniques that could be useful during the Hackathon (web scraping, text mining, geocoding, etc.). This preparation phase was well received by the participants and the organising team received positive feedback.

c. Transforming this event into an experiment

Since the Hackathon, a small team (including Census department members, SSP Lab members, IT members and some participants in the Hackathon) has been working to the development of a prototype, implementing some ideas that emerged during the Hackathon and adding some new

functionalities. This is still a work in progress.

4.2 Detecting wages/paid hours anomalies in employer payroll declaration statistical databases

Sponsor: INSEE Social Studies Directorate (Employment and Professional Income unit);

Team: SSP Lab (1 member), Statistical Methods Unit (1 member), Employment and professional income unit (2 members);

Schedule: from January to December 2018;

Expected deliverables: experimentation report, guidelines for implementation, and methodological and academic contributions.

a. A major change in the employer payroll and social contribution declaration format offering new opportunities

The Annual Declaration of Social Data ("déclaration annuelle de données sociales", DADS), mandatory fulfilled each year by each employer and to which reported individual wage-earner information is transmitted to fiscal and social services for payroll and tax purposes as well as for calculating social security wage-earners rights (e.g., pensions), has been replaced since 2016 by a monthly Nominative Social Declaration information. This change of sources completely modifies the national statistical service of information on employment and wages that relies on, but also provides the opportunity to rethink the automatic anomaly detection process implemented in the statistical production line, as the latter is deeply modified to integrate these new data. An adapted automatic detection of such anomalies would lead to productivity gains in the subsequent editing procedure.

b. Machine learning contributions to anomaly detection of wages/paid hours data

The experimental project carried out with the department of Employment and Professional income of the Social Studies Directorate tests different machine learning-based algorithms for anomaly detection of net and gross wages and related paid hours. The project has so far investigated unsupervised algorithms, such as fuzzy association rules, isolation forests and local outlier factors on a small scale, with the intention to provide a probability score of an outlying position. The next steps will be to constitute a sample of observations labelled as anomalies or not to evaluate the performance of the methods tested in comparison to the current one (based on

a classic regression) and to apply the methods selected on larger scales.

4.3 Mobile phone data

Sponsor: INSEE Regional Studies Directorate;

Team: SSP Lab (one permanent member and one intern);

Schedule: several experimentations since 2016;

Deliverables: institutional and academic contributions, data-processing techniques and experimental prototypes.

Mobile phone data has proven to form an exciting new data source for official statistics. The SSP Lab intends to explore the institutional, legal, technical and methodological challenges that come with the integration of mobile phone data in official statistics.

a. Data access at Orange Labs

Access to a pseudo-anonymised dataset collected by Orange for billing purposes has been made possible through an agreement between Orange Labs, Eurostat and INSEE. The dataset consists of Call Detail Records (CDR) describing information on each phone call and text message (SMS) sent or received by Orange users in the period from May to mid-October 2007.

b. First experiments

A number of experiments have been performed. The goal of the first experiment was to detect urban zones thanks to mobile phone data and application of supervised classifiers (Vanhoof, Combes, de Bellefon, 2017). The second experiment intended to measure the residential population by using the CDR during nights and advanced treatments of the data (de Bellefon, Givord, Sakarovitch, Vanhoff, 2018). The last experiment is still in progress and analyses the segregation by combining mobile data and fiscal data (Galiana, Sakarovitch, Smoreda, 2018 presented in this conference).

c. Future prospects

These experiments showed that mobile data are extremely rich, but could be unsuitable for some applications because of location imprecision or representative bias. In order to exploit the whole

richness of information of these data, the following experiments will focus on the estimation of population present within a given place and time (as opposed to the residential population). Different time and geographical scales will be explored, potentially with signalling data. A new agreement for continuing the collaboration is ongoing.

5. Concluding remarks

Although the Lab is still in its early days, it is clear that the opportunities for new synergies with production units have been positively welcomed throughout the official statistics service ecosystem. The creation of the Lab is proving to be promising in terms of acquiring, maintaining and disseminating data science knowledge. What's more, it is already mobilising people within our organisation, and helping to inspire and motivate the next generation of our statisticians.

Bibliography

- Combes, S and Bortoli, C: Apports de Google Trends pour prévoir la conjoncture : des pistes limitées, *Note de conjoncture*, March 2015
- Combes, S., Bortoli, C. and Renault, T.:" Nowcasting payroll employment with traditional media content", New Techniques and Technologies for Statistics conference, March 2017.
- de Bellefon, M.-P., Givord, P., Sakarovitch, B., Vanhoff, M. : "Allô, où es-tu ? Estimer la population résidente à partir de données de téléphonie mobile, une première exploration, in revision in *Economics and Statistics*, 2018
- Galiana, L. and Sakarovitch, B. and Smoreda, Z.: "Understanding socio-spatial segregation in French cities with mobile phone data", 2018, presented at this conference.
- Vanhoof M., Combes S., de Bellefon M.-P.: "Mining Mobile Phone to recognize Urban Areas" in A. Petrucci and R. Verde, eds., SIS 2017 Statistics and Data Science: New Challenges, New Generations. Proceedings of the Conference of the Italian Statistical Society (Florence: Firenze University Press, 2017) 1005-1012.

Annex: Ongoing projects

Exploring new data

- Student evaluation log files: more information to assess student response strategies [with the Education ministry statistical service, and IT department of INSEE]

- Mobile phone data - residential population, social segregation [with Orange Labs, ESSnet Big data] (cf 4.3)
- Branch agreements: What protection for employees? A textmining approach [with master's students]
- Satellite Data and city heat islands [with the INSEE geographic methods unit]
- New panel data about professional careers of both wage-earners and entrepreneurs [with INSEE social studies Directorate]

Applying new methods

- *Les champs de Sirene*: Automatic detection of employer in census [with IT department and hackathon participants] (cf 4.1)
- Detecting wage anomalies in employers' payroll declarations statistical databases [with INSEE Social studies Directorate] (cf 4.2)
- Machine learning for predicting careers and wages for microsimulation models [with INSEE Economic Studies Directorate and IPP-Paris School of Economics]
- Peer effects in Education [with the Education ministry statistical service]

Ongoing methodological reports

- Decomposition methods for inequality analysis
- Selection on observables: Propensity score in R (just released)
- Duration models in statistical studies

Other studies in collaboration

- Constructing control group for poor city districts (INSEE Regional studies)
- Differentials in peer effects in high school success (Education)
- Quantifying the effect of school avoidance on segregation (Education)
- Heterogeneity of the performance of high schools in France (INSEE Economic studies)
- Firm role in gender wage gap (INSEE Social studies)
- Wage discrimination against descendants of immigrants (INSEE Social studies)
- Evaluation of the 2014 Unemployment Insurance Agreement (INSEE Economic studies, Acoss)

Training sessions given in 2018

- Machine Learning (2d)
- Textual Analysis (1d)

- Python for the data science (in preparation)
- Decomposition methods for inequality analysis
- Evaluation of public policy

Dissemination and collaborative work practices

- Hackathons, collaborative workshops,
- Reading groups: Machine Learning and econometrics for career analysis with panel data (in preparation)

Dissemination

- Intranet, extranet SSM, Yammer, blog, Github
- Big Data newsletter, Big Data seminars

Networks

- Eurostat - Big Data task force, Essnet BigData I (and II)
- Eurostat - Grant City data (mobility and phone data)
- Academics - CREST, IPP-PSE, Dauphine