Eurostat: Stats in a Wrap

Innovative approaches in statistics – Part I: Inventiveness and AI

Release date: 18 July 2024

SPEAKERS
Albrecht Wirthmann (Eurostat), Nikos Roubanis (Eurostat), Frankie Kay (Central Statistics Office of Ireland), Jonathan Elliott (host)

Jonathan Elliott
Stats in a Wrap, the podcast series from Eurostat.

Jonathan Elliott
Welcome to another edition of Stats in a Wrap, the podcast all about statistics from Eurostat, the statistical office of the European Union. In the wrap café we like our data tasty and piping hot, and we don't like waiting too long for it either. People want official statistics faster these days, and they want them to be easier to access.

We are in the click society - instant, high-quality information and as much as we want, is regarded as, well, practically a human right. Eurostat's six quality assurance principles are, let us not forget: relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability and coherence. But each principle cannot be allowed to diminish any of the others.

To deliver good stats, the methods used to achieve those standards need to be constantly updated. In the wrap café today, we're talking about innovation. That's right, inventing stuff. This time, it's the inventive recipes used to make official statistics. And the people who are doing the inventing are the members of the European Statistical System - that's the family of stats organisations in the EU and the European Free Trade Area (or EFTA).

So, we have two special editions, each showcasing a different aspect of the innovation: inventiveness and experimentation behind the scenes of the ESS innovation agenda, as it's called. To help guide us through this fascinating subject we have again in the wrap café an old friend of the podcast, Mr. Innovation himself at Eurostat, the head of the unit responsible for methodology and innovation, Albrecht Wirthmann. Albrecht, welcome.

Albrecht Wirthmann
Hi, Jonathan. Nice to be here again.

Jonathan Elliott
Lovely to have you here again. And his colleague, Frankie Kay. She's the Chief Information Officer at the Central Statistics Office of Ireland, and one of her responsibilities is the newly launched one-stop shop for artificial intelligence and machine learning for official statistics. Welcome, Frankie.

Frankie Kay
Hi, Jonathan. Thanks very much for having me. It's lovely to be here.

Jonathan Elliott
I understand you're looking for a snappier title than - here we go - 'one-stop shop' for artificial intelligence and machine learning for official statistics.

Frankie Kay
Absolutely! So, if anybody's got any bright ideas…I probably should be using ChatGPT to come up with a new snappy name.

Jonathan Elliott
No, no, we've got a band of loyal listeners. So, if anybody can put forward suggestions for a better name than the 'one-stop shop for artificial intelligence and machine learning for official statistics', we'd love to hear them. Finally, we have Nikos Roubanis.

As he heads up Eurostat's transport statistics unit, he'll be talking us through an amazing project to dramatically improve the way stats on maritime shipping are collected. Ingenious does not come close to describing it. So, we're looking forward to that. Welcome, Nikos.

Nikos Roubanis
Hello. Thank you for inviting me.

Jonathan Elliott
Thank you for being here. Well, let's just dive in with some sort of general introductions from Albrecht. If you could just talk us through a few essential things here. Why is innovation such a big thing in statistics, and why is it important?

You very kindly joined us last February on Stats in a Wrap to talk about experimental statistics, and you said there was a strong need among stakeholders for timely production. This was in the wake of the Covid-19 pandemic. So, I'm just wondering what has happened since then, and are there any further factors at play in innovation? Can you update us?

Albrecht Wirthmann
Of course, experimental statistics is part of the innovation agenda of the European Statistical System. So, it's the output of the innovation that we are finally publishing. What has happened since then, apart from the crises happening, the topic of artificial intelligence is discussed everywhere.

There we've seen a real hype in the use of artificial intelligence for different purposes, including official statistics: we are responding to these developments to stay on top and to enable the future innovation in the production and publication of statistics.

Jonathan Elliott
Just wanted to see if you could talk me through one thing, which is the extent to which innovation is enabled by opportunity. In other words: if there's more data out there, does that mean you can more readily innovate because it's available and you can start doing things with it? So, is it one of those

things where you have to be a bit of an opportunist, a bit of an entrepreneur, to keep an eye open for new data sources so that you can then go on and do new things with them.

Albrecht Wirthmann
We have to take the opportunities that are out there. We've prepared now for a legislation that allows the use of these datasets, so we are exploring different opportunities to use these data and technologies to better fit the needs of users for statistics, and to develop innovative statistical products.

Jonathan Elliott
What's exciting about this is that it's very much kind of a cooperative network. You're almost like a college system in a university, and you've got lots of different departments all working away, and you're all trying to coordinate everybody to work together optimally. Is that right? Is that how it's working to get a sort of synergy going?

Albrecht Wirthmann
One of the main purposes of the innovation agenda is to make those innovative developments visible inside the statistical offices and between statistical offices. And also, to make them known to stakeholders, such as the universities, the academia, for example, but also for enterprises and users of statistics, such as politicians. In this way we are trying to create synergies and spread innovative ideas that can be taken up by the statistical community.

Jonathan Elliott
And this sounds like it's in the best traditions of academic computer science, if you like, open source, everyone sharing, everyone collaborating, no market forces hoarding intellectual property in a vault. Everyone's being very open about things to try and get the best innovation going. I mean, am I overidealizing it?

Albrecht Wirthmann
This is the part of our strategy - to be open in the development of methodology and processes: meaning, to reach out to other stakeholders. In this way, we are also creating opportunities, for example for enterprises. We are active in the data market. So, we are partially operating now into this new data market. Through this, we are hoping to contribute to the development of this market as well.

Person on Street 1
I think artificial intelligence could be good for, like, polls and asking the right questions and maybe getting more complicated answers out of people.

Person on street 2
I think artificial intelligence can be used in almost every field that is imaginable. It has to be some kind of specialised for it. I'm not into the topic that much that I can tell you how it works or something like that, but I believe if you do it correctly, artificial intelligence can learn, and it can improve every field of study or work.

Jonathan Elliott
Well, it won't surprise anybody that when we're talking about innovation and statistics and statistical methods, artificial intelligence is never far away. Frankie Kay is the lead on AI in statistics for this

network, for the ESS. Can you just tell us a bit about what AI can do in official stats? Why is it so powerful, and why can it be a game changer?

Frankie Kay
I guess, I suppose it's like almost any sector at the moment, AI has the potential to transform it. And I would say the introduction of ChatGPT seems to have caused a real acceleration in the interest in the use of AI, including in official statistics. And I suppose there are many different challenges in official statistics. We're trying to get good quality data and information out. We need to do it in a timely manner. We need to be able to do it at a granular level.

It needs to be relevant. And that all comes down to one fundamental thing, which is having good, high-quality data that we can then apply those methods that Albrecht was describing, applying standards on top of that, so that we can produce the information that people are interested in. And I suppose a concrete example: we in Ireland, we still conduct a census, and we ask people what their occupation is.

And if you think about the types of occupations there are at the moment - they are very different to those from 10, 20 years ago. So, people write that in, and we can then use artificial intelligence to say what new roles are emerging. An example at the moment might be something called prompt engineering, which is how you ask ChatGPT really relevant and precise questions, and that might be something that people now add onto a census form that we wouldn't have had in our list.

So, AI and machine learning can then help create those different types of occupations, and we can see which new ones are emerging, rather than us literally having to go through all of those manually. So that would be a really concrete example of where we're seeing AI machine learning. And the large data that we're looking to deal with, if you think about prices as another example where we're trying to figure out inflation.

And traditionally people would literally go around supermarkets and write down the prices of lots of different products. When shopping, we get the scanners, and we can take that scanner data that can provide us information about what prices are and how much people are buying what particular goods. And we can't really make sense of that type of data without using new technologies such as machine learning and AI.

Jonathan Elliott
Just on the way that AI works with job descriptions. I mean, it's amazing. It actually goes out there and reads job ads and collects descriptions in a human like way. It's called, I understand, 'web scraping', which sounds rather unpleasant and painful, but can you just tell us what this is - web scraping - and what the AI does to get these job descriptions?

Frankie Kay
You have little, almost, well, they're not robots, but they’re almost like little spiders that go into multiple different websites, and effectively look to see all the different jobs that are being advertised. But then if you think of a job, it's probably on more than one website, potentially. So, what you then need to do is try and understand whether that's actually a different job or whether it's the same job.

So it's not just a matter of taking multiple scans of all these different pages of online job portals, you also then need to bring them together, categorise them, understand if they are repeats, and therefore

come up with a clear view in terms of how many vacancies we think there are, and I suppose also things like what type of skills are people looking for, what types of roles…So again, a really broad range of information.

Jonathan Elliott
Fantastic. Now, if you don't mind, introduce us to the one-stop shop. It has a rather appealing notion that you can go into a single place and get all your AI needs met under one roof, as it were, very appealing idea. Just tell us a bit about that. I mean, it's only just launched, isn't it? It was this April.

Frankie Kay
It's one of the biggest projects, I believe, that Eurostat have launched. One of the things in statistics, again, Albrecht was referring to this, is a need to have some commonality in how we produce the statistics, so that everybody understands that we're all measuring things in a similar way, so that we can compare. And also, we have limited resources in NSI's. So therefore, we don't want to in multiple different places effectively be doing the same thing when we can share.

Because, as you mentioned, Jonathan, we're not under commercial pressures. We are not fighting each other. So being able to share the work that we do is something we are very proud of, and as the use of MI - machine learning - increases, there is a need to try and standardise where we can, try and understand best practice.

So again, rather than every individual country attempting to do that, we wanted a more structured way that we would provide somewhere where anybody working in an NSI - or indeed, anybody that's interested, it's not just about national statistical institutes - can come and look about what we're doing.

So, it's about helping provide best practice, it's about helping provide training. It's also looking at the underlying models that we use in AI to see can they be standardised, and if they can, to what degree that we can reuse what we're doing in those models across many different countries. And therefore, that's what this is about, it's providing that central place of resources that we can learn from each other, but also that we can share.

Jonathan Elliott
Yes, and I can imagine that having consistent methods across all the Member States is vital for Eurostat, because it needs to make sure that the data collected by AI is done in the same way across all the different Member States. You can't have everybody doing their own little kind of homemade version of AI.

Frankie Kay
No, you can't. Fundamentally, the allocation of EU budgets across Member States is based on economic statistics that are produced by individual countries. So, it's absolutely vital that they are done in a consistent way to a good level of quality across all Member States.

Person on street 4
Do I trust the data that's collected by AI? It's a bit the same as trusting what you read on the internet. Because - I don't always believe it. Because I like to check sometimes with my own experience, or maybe the experiences of colleagues and friends.

Person on street 5
I trust the data collected by artificial intelligence because it is using so much data, it is a very broad spectrum of data that it is able to scrape and scour and look for. So, I agree that I would trust it. Yes.

Jonathan Elliott
We're going to talk now about the lighthouse project. Now this is a collection of promising innovations in development at the ESS. One of them has accelerated statistics on maritime traffic coming and going from the commercial ports of Europe in the Mediterranean, Atlantic and North Sea. And Nikos Roubanis is here to guide us through that project. Nikos, just tell us first how these stats are conventionally collected. It's not a speedy process, is it?

Nikos Roubanis
So, the way we collect and publish statistics is that ships arrive in EU ports, they have to declare their arrival, they have to declare where they come from. They have to declare what they carry and how much the weight of the goods they carry. So, it is a very detailed reporting system, and basically it takes a lot of time, and we publish this with a delay of 12 months while everybody knows at every moment where the ships are.

So, the idea behind this project is to use real-time data to actually extract similar statistics - like the statistics we receive - but much faster. So, at the end of each quarter, we can be in the position to say exactly how many ships arrived into EU ports. And I'm talking about ships that have some economic activity. So they are loading on or unloading goods, because basically this is the indication of economic activity that we are looking for.

Jonathan Elliott
Now, the real-time data that you get comes from a unique source. It comes from a very particular and somewhat unexpected one. Can you just talk us through that?

Nikos Roubanis
Yes, basically each ship, while navigating, is sending a signal; a signal saying what ship it is, the type of ship and where it is. So, this is called Automatic Identification System (AIS). It is captured by stations which are at the coast, and satellites.

Now these signals are monitored constantly by the European Maritime Safety Agency for safety reasons and can be used for statistics. For us, the big question is: how do we relate these signals - and we explore the data which are in the signals - with our data, with the statistics that we collect.

Jonathan Elliott
But somebody in the transport unit must have thought: ah, EMSA have got all this amazing data, did you realize that? And suddenly, a light bulb moment came on. Did you think - we could use that! That could have some real applications for us. I mean, it shows very clever lateral thinking, if I may say, or maybe these things, maybe data sources are constantly reviewed and analysed by people at Eurostat. How did this idea originate?

Nikos Roubanis
Well, I have to say it is an innovative data source, because we are bound to produce the statistics which are produced at national level. So, our first reaction would be to ask Member States to work

faster, send us the data more quickly and publish a little bit faster, but we would never be able to aggregate all this information from 27 countries just a few…a couple of weeks, maybe, after the end of each reporting period.

So, we had to look for alternatives, and in fact, EMSA can provide also some assurance that this information is relatively complete. So, then the next problem would be: and how do we actually make statistics out of new signals? That was the real challenge for us.

Jonathan Elliott
But you have to come up with a system that's more than just a nice, clever theory or an experiment. It actually has to work. It has to deliver a real use case, solid stats that are going to be useful continually and reliably. So just tell us a bit about that project in itself. Getting that data to perform for you was not a straightforward process, was it?

Nikos Roubanis
We looked at data in the past, aggregated AIS signals in the period of 2015 to 2019, so pre-Covid period statistics. We analysed the various fields of the signal records and compared the results with statistics received for that period.

Then we aggregated the AIS signals by vessel type because the vessel type is a parameter in the AIS signal record. Then we tried to match port calls - the number of ships arriving to ports (that is the meaning of port calls) - by vessel type, between the AIS signals and statistics we collected over that period in the past, so, 2015 to 2019.

We identified many challenges, for example, matching the position of the vessels from the AIS signals in the same way as in statistics received and identified, for example, stationary ships or ships without economic activity, ships that are not loading or unloading goods or passengers. Sometimes we had even to contact data providers, so ports or national statistical authorities, to clarify eventual outliers and differences in the two data sources.

And the objective of this comparison was to create a method, an algorithm, of deriving the statistics received later from the instantaneous, aggregated AIS signals, and benchmark this method with data received in this past period.

Jonathan Elliott
Presumably, now you've got a robust platform where everything is functioning, and you can actually operate this now and it's delivering the goods. Is that true? I mean, has it matured?

Nikos Roubanis
The best match between the two data sources, AIS and past statistics, was achieved by adjusting past statistics with current trends from AIS signals. So, we observed a very good match between the statistics and trends from AIS data. We still need to do some further testing, however, to make sure that the algorithm can perform well, even in exceptional situations. Results and methods are published currently as experimental statistics.

Now, our intention is to regularly update these experimental statistics and when we will be fully convinced that the predictions for the present period are always accurate, we will pass into regular statistical production: maritime statistics from AIS data.

Jonathan Elliott
So, beta testing at the moment, but quite promising. One of the things about innovation in tech is always scalability, that it can be applied elsewhere. I was just wondering whether other statistical organisations have been in touch. Are you aware of other countries, for example, or other geographies around the world using this? Because it seems to me it has global applications.

Nikos Roubanis
Many European countries are already doing maritime statistics from AIS. Each country is doing it a little bit differently. What is the value of us is: first, we have the data which are readily available. The data are available at an EU agency, so we can have access to the data. And second, our method is a European method. It can be applied probably to every country. And that is our next step - not to scale up, but scale down and produce national statistics.

Frankie Kay
If you don't mind, perhaps just very briefly coming in there. First of all, I mean, Ireland is a country that has also been producing and looking at this as well, we're using AIS in a similar way to Nikos but to go back to the…to briefly what I was saying about the one-stop shop, this is exactly what we mean. It's about having standard methods.

It's about us not all doing the same work, because I know the UK previously, where I used to work, they've also done work in this area, the UN have. So, this is a prime example of producing those methods and standards for the good of everybody. So that's what the one-stop shop is trying to do, is to gather all of those and make them available for people.

Jonathan Elliott
Yes, well, suddenly within Europe, but maybe I'm getting overexcited with too much coffee here. But shouldn't there be a global one-stop shop, a global network of statistical institutes all sharing on this because it's applicable anywhere, isn't it?

Frankie Kay
There is a UN platform where similar types of data is available, and I work quite closely in the one-stop shop, we were liaising with the UN as well for exactly what you're saying. So no, you're not having too much coffee, it's exactly what we should be doing.

Jonathan Elliott
Yes, I mean, transport seems particularly blessed with large quantities of, well, sensor data, for example, from road toll plazas, perhaps aviation, rail, there are plenty of others. I mean, Nikos, perhaps you could just tell us about some of the other areas in your unit where you're looking at data from unusual sources as potential raw material for statistics.

Nikos Roubanis
Transport is a very good case to showcase the use of innovative data. Everything that moves today generates information, and this information can also be used for statistics. Similarly to ships, we also

know the position of rail vehicles, each vehicle, and we can identify also which are the vehicles, which transport goods and people.

So that is a next project that we are planning to pursue, to try to replicate and create new statistics on railways, which are very timely. Road transport is also very, very much connected. People are connected. Vehicles are connected. Vehicles are sending diagnostics data in real time. So that is also another domain where we plan to develop methodologies and ways to produce interesting statistics, and indeed, we have also a project on road traffic and mobility.

Jonathan Elliott
Fantastic! Well, we've run out of time, and it only remains for me to thank everybody for their amazing contributions on this first of two podcasts about innovation in statistics in the European Statistical System. So, I'd like to say thank you very much to Albrecht Wirthmann as the Head of Unit 'Methodology, innovation and official statistics' at Eurostat. Albrecht, thank you for joining us.

Albrecht Wirthmann
It was a pleasure, thank you.

Jonathan Elliott
And Frankie Kay, the Chief Information Officer at the Central Statistics Office Ireland and also leading the one-stop shop for AI and machine learning in official statistics. Frankie, thank you very much for joining us today.

Frankie Kay
It's been a pleasure. Thank you.

Jonathan Elliott
And finally, Nikos Roubanis, Head of Unit for transport statistics at Eurostat, thank you so much for sharing your project with us. Thank you very much, Nikos.

Nikos Roubanis
I thank you too.

Jonathan Elliott
If you've enjoyed Stats in a Wrap, don't forget to follow us on social media and share our adventures with friends and colleagues, where the show can be found on Spotify, Apple and all the usual places. And of course, join us for the next innovation edition in September, when we'll be dishing up more flavoursome insights from Eurostat. This time about how chatbots are helping statisticians and the public get their statistics in Norway. Join us then, but for now. Goodbye!