En Clave INE

Episodio 3. Nuevas fuentes de datos y estadísticas experimentales

Con Alfredo Cristóbal Cristóbal
Director General de Estadísticas Económicas del INE

Suelí Lorenzo

Bienvenidos a En Clave INE, el podcast del Instituto Nacional de Estadística, un espacio para entender mejor la realidad que nos rodea a través de las estadísticas oficiales. En cada episodio hablaremos con nuestros expertos para conocer las estadísticas, para descubrir qué hay detrás de los datos y para mostrar su utilidad en la vida cotidiana. Soy Suelí Lorenzo, muchas gracias por acompañarnos en este episodio.

Cuando pensamos en el INE nos vienen a la cabeza los censos, las encuestas, los datos, pero detrás de cada cifra hay mucho más. Hay métodos estadísticos, hay tecnologías y hay profesionales que trabajan para reflejar una sociedad en constante transformación. Esto nos lleva a una pregunta, ¿cómo puede un organismo como el INE seguir midiendo la realidad que está en un cambio constante?

Hoy estamos aquí para hablar de innovación, para hablar de nuevas fuentes de datos y de estadísticas experimentales y para eso contamos con Alfredo Cristóbal, Director General de Estadísticas Económicas del INE.

Bienvenido, Alfredo.

Alfredo Cristóbal

Gracias, Suelí, me encantó conversar contigo.

Suelí Lorenzo

Alfredo, cuando hablamos de innovación normalmente pensamos en empresas tecnológicas, en nuevos dispositivos, en inteligencia artificial, pero también hay innovación en la estadística oficial.

¿Qué significa innovar en un organismo como el INE?

Alfredo Cristóbal

De manera sintética te diría que innovar es mejorar lo que ya hacemos con el objetivo de crear valor. Es decir, en estadística podríamos pensar en producir información más útil, más rápida y de forma más eficiente generando menos coste.

Suelí Lorenzo

Durante mucho tiempo hemos asociado la estadística oficial a cuestionarios que rellenan los ciudadanos y las empresas.

¿Sigue siendo esa la base del trabajo del INE?

Alfredo Cristóbal

No, tradicionalmente es cierto, como bien has dicho, que la producción del INE se ha basado en lo que llamábamos la producción basada en encuestas. Es decir, nos acercábamos al ciudadano, a la empresa y le preguntábamos determinadas cuestiones que nos servían para producir luego las estadísticas de un determinado sector. Lógicamente esto ha mudado y desde hace varios años nosotros estamos basando más nuestra producción en los registros administrativos.

Los registros administrativos tienen la ventaja de que es una información que previamente ya se ha recogido de la población y que por lo tanto no necesitamos volver a contactar con ellos para preguntársela. Por ejemplo, cuando una persona se da de alta para trabajar en la seguridad social tenemos que facilitar muchísimos datos que pueden ser utilizados posteriormente a efectos estadísticos o como cuando rellenamos el impuesto sobre la renta, por ejemplo, para la agencia tributaria, de la misma manera generamos también una cantidad de información que se la transmitimos a la agencia tributaria pero que también podríamos luego utilizarla para los fines del INE. Más en el día de hoy te diría que no solamente encuestas o registros administrativos están siendo actualmente las formas en las cuales nosotros nos basamos para producir estadísticas.

A día de hoy también son importante también la información de las bases de datos privadas. Las bases de datos privadas son un gran cúmulo de información que nosotros también podremos utilizar y podemos y estamos haciéndolo para generar información estadística. ¿Cuál es la clave? ¿Cuál es la clave del éxito? La clave del éxito va a ser integrar toda esa información basada en encuestas, basada en registros y basada en datos privados de forma que podamos obtener la estadística más eficiente y más ágil.

Suelí Lorenzo

Y para entender mejor este cambio, esta transición de esta manera más tradicional de elaborar las estadísticas a esta incorporación de estas nuevas fuentes, hay un momento que sin duda fue decisivo que es la pandemia. El COVID-19 pues obligó a muchas empresas pues a adaptarse rápidamente para buscar nuevas formas y nuevas fuentes de información.

¿Qué cambió en la manera de trabajar en el INE a raíz de esta pandemia?

Alfredo Cristóbal

Tienes toda la razón, Suelí. La pandemia del COVID marcó un antes y un después en lo que es la producción estadística. Realmente estábamos confinados hacia mitad del mes de marzo del año 2020 y durante unas semanas fuimos poco capaces de producir la información que la sociedad necesitaba con la rapidez que se necesitaba en aquel momento. La sociedad estaba cambiando día a día y no solamente me estoy refiriendo a información de tipo sanitario, que sí, sino también información de tipo económico.

Hay establecimientos que estaban cerrados por decreto, otros que cerraron porque no tenían negocio, otros que estaban abiertos y por lo tanto era imperativo conocer cómo se iba desarrollando la economía en aquel momento. Por eso y por vez primera en el INE fuimos capaces de analizar las ventas diarias de las grandes empresas de la agencia tributaria que nos arrojó algo de luz en ese sentido. Pero también, y te pongo un segundo ejemplo, fuimos capaces de realizar de forma experimental un análisis de la movilidad de la población basada en la información de los teléfonos móviles, es decir, del posicionamiento que tenían estos teléfonos móviles y de cómo la población se movía entre áreas, sin ninguna necesidad de conocer la identidad de los tenedores de los terminales, es decir, garantizando en todo momento la privacidad de la información.

De esta manera pudimos conocer cómo transitaba la población entre áreas y que, de otra manera, haciéndolo a través de una estadística tradicional, pues hubiera sido, si no inviable, por lo menos costosísima.

Suelí Lorenzo

A la vista de lo que comentas, entiendo que la cuestión de fondo ante este cambio de paradigma y ante esta realidad en constante transformación es si una estadística puede permitirse años para transformarse, para evolucionar.

Alfredo Cristóbal

No, evidentemente no.

La sociedad, la economía, las formas de producir cambian constantemente y el INE pues tiene que adaptarse de la forma más rápidamente posible. Por ello, por ejemplo, te citaré un ejemplo. Una forma tradicional que tenían los turistas como solución habitacional de alojamiento eran los hoteles, pero, sin embargo, desde hace unos años a esta parte, hemos visto que ha surgido otro fenómeno importante que es el fenómeno de las viviendas turísticas.

Las viviendas turísticas facilitan una solución alternativa a los hoteles y de una forma competitiva o incluso mejor desde el punto de vista de los precios, con lo cual es elegida por muchísimas personas como alternativa. Y, por lo tanto, nosotros ahora tenemos que ser capaces de poder medir ese fenómeno y hemos utilizado también otro tipo de herramientas innovativas que es extraer la información de las plataformas donde se anuncian estas viviendas para poder dar una síntesis, un perfil de cómo están afectando al turismo y a la economía las viviendas turísticas.

Suelí Lorenzo

Bueno, suena realmente muy prometedor y como fuente de información, pero me imagino que también tendrá sus dificultades.

Alfredo Cristóbal

Por supuesto, y no pocas. Realmente uno puede imaginar que las viviendas turísticas lo pueden ser a tiempo completo, es decir, todo el año puede ser vivienda turística, puede ser una vivienda de temporada, puede ser simplemente una solución compartida, es decir, una vivienda principal en la que el propietario alquile una o varias habitaciones para uso turístico. Por ejemplo, distinguir entre esos tipos es complejo. Pero también, naturalmente, tenemos que pensar que no todas las viviendas se anuncian en plataformas y que, con el objetivo de maximizar la posibilidad de poder alquilarla, un propietario puede anunciar su vivienda en más de una plataforma. Con lo cual, todo ese tipo de limitaciones tenemos que ser capaces metodológicamente de poder salvarlas.

Suelí Lorenzo

Alfredo, venimos hablando de la incorporación de nuevas fuentes.

Nos acabas de comentar el caso de las plataformas turísticas. ¿Podrías generalizar y darnos una serie de razones por las que el INE debe de trabajar con estas fuentes?

Alfredo Cristóbal

Te voy a citar tres razones fundamentales. La primera es la rapidez. La forma de gestionar esta información nos proporciona información muy rápida que la sociedad necesita para su toma de decisiones.

En segundo lugar, te diré el grado de detalle o la granularidad que se puede comentar desde la perspectiva estadística. Naturalmente, los ciudadanos entienden y son muy necesarios los indicadores macro, pero cada día lo que necesitan es conocer ya no solamente cómo evoluciona la nación o su comunidad autónoma en determinado fenómeno, sino conocer qué es lo que está pasando en

su barrio. Es decir, necesitamos lo que se denomina acercar la estadística al ciudadano. De esta manera, necesitamos fuentes y procedimientos que nos permitan acercar esa estadística para que el ciudadano conozca qué es lo que está pasando en su entorno.

Y, en tercer lugar, te diría la eficiencia. Realmente, de esta manera, somos capaces de producir mucho más rápidamente y mucho más granularmente, como he citado antes, pero también con un menor coste. Y, por lo tanto, digamos, esas serían las razones fundamentales de la utilización de estas fuentes.

Suelí Lorenzo

Si te parece, también detengámonos en alguna de estas nuevas fuentes. Respecto al uso de datos de internet sobre *web scraping*, ¿podrías contarnos qué es exactamente y cómo se utiliza en estadística?

Alfredo Cristóbal

Sí. Vamos a ver. El *web scraping* es una técnica para recoger datos directamente de internet cuando éstos están disponibles públicamente.

Es decir, te voy a poner un ejemplo que se va a entender fácilmente. Cuando vamos a un hotel, nosotros previamente lo que hacemos es entrar en su página web, que nos dice todos los tipos de habitaciones que tiene disponibles y a qué precio, y elegimos la que más nos conviene. Pues imaginemos un procedimiento automático por el cual una especie de robot pudiera ser capaz de entrar en cientos, en miles de páginas web y obtener todos los listados de todas las habitaciones disponibles a todos los precios, cada uno de ellos.

De esta manera, esta información se podría obtener, se podría recoger directamente para utilizarlo en el procedimiento estadístico sin necesidad de tener uno o varios agentes que tuvieran que ir a los hoteles a tomar estos precios. Ese es el *web scraping*. Es útil en este caso, por ejemplo, para el índice de precios turísticos o para el IPC, pero también se puede utilizar en otros ámbitos de la estadística.

Suelí Lorenzo

Alfredo, anteriormente nos hablabas del uso que habíais hecho de la información agregada, de los datos de telefonía móvil a raíz de la pandemia. Es una fuente muy interesante porque sin duda genera una gran cantidad de información y nos gustaría saber qué aportan estos datos a la estadística oficial.

Alfredo Cristóbal

Muy fácilmente te hablaré de qué responde a las tres ventajas de las que he citado en una pregunta anterior, porque por un lado es información muy rápida, es decir,

los datos se pueden obtener de las plataformas, perdón, de los operadores móviles de una forma muy rápida, aportan muchísima granularidad y ahora citaré un ejemplo, y además realmente son muy eficientes porque reducen muchísimo la carga a los informantes.

Por decirlo de otra manera, hacer una encuesta para poder gestionar esa misma cantidad de información sería inviable o tendría un coste elevadísimo. Pero por el tema de la granularidad te voy a citar un ejemplo que se va a entender perfectamente. En el estudio experimental sobre movilidad que hicimos en 2020 fuimos capaces de ofrecer, por ejemplo, a nivel agregado la información de, vamos a poner por ejemplo Madrid, las personas que viven en un determinado barrio, como puede ser el barrio de Vallehermoso, y que el 15 de agosto estaban en el área de la playa de Levante de Benidorm.

Es decir, pudimos llegar hasta ese nivel de granularidad sin necesidad de tener la información privada de cada uno de ellos, sino simplemente tratándolos de forma agregada. Es decir, la granularidad que nos puede ofrecer esa información es tremenda.

Suelí Lorenzo

Un pequeño detalle, porque en tu respuesta acabas de comentar, has hecho referencia a la carga a los informantes, pero podrías explicar un poco mejor qué entendemos por carga en el ámbito estadístico.

Alfredo Cristóbal

Sí, la carga es el tiempo, el esfuerzo, podríamos decirlo mejor, que tanto los ciudadanos como las empresas tienen que dedicarnos para ofrecernos la información que nosotros les estamos solicitando. Si naturalmente podemos reducirlo, eso irá en beneficio tanto de ellos como naturalmente de nosotros.

Suelí Lorenzo

O sea, que innovar también de alguna manera es reducir ese esfuerzo que pedimos a ciudadanos y a empresas.

Alfredo Cristóbal

Completamente.

Suelí Lorenzo

Qué interesante, Alfredo. Yo creo que esto cambia la imagen tradicional que los ciudadanos podemos tener de la estadística oficial, con la incorporación de todas estas nuevas fuentes de datos. Pero yo creo que esto también nos lleva a una cuestión importante o delicada. Si los datos ya no proceden tanto de las preguntas

que directamente se hacen a los ciudadanos y a las empresas en las encuestas, ¿cómo podemos estar seguros de que el dato es fiable?

Alfredo Cristóbal

Y es una pregunta muy pertinente, Suelí. Realmente los datos por sí mismos no garantizan la fiabilidad de los resultados estadísticos.

Es justamente el tratamiento metodológico que se haga posterior con ellos. Por ejemplo, en el caso de la telefonía móvil, podría darse el caso de personas que viajan con varios terminales móviles, con un teléfono, dos teléfonos, dos teléfonos y una tableta, por ejemplo. O podría darse el caso de una familia que comparte determinados dispositivos, o simplemente una persona que ha comprado una tarjeta prepago en otro país diferente.

Por lo tanto, tenemos diferentes problemas metodológicos que tenemos que identificarlos, en primer lugar, y segundo, solucionarlos a través de aplicar la metodología con los datos que nos proporciona el operador móvil para poder garantizar la calidad del resultado.

Suelí Lorenzo

Es decir, que incorporar una nueva fuente de datos no significa simplemente agregar datos, sino que entiendo que detrás, como nos cuentas, hay un enorme trabajo metodológico.

Alfredo Cristóbal

Efectivamente. Realmente es un trabajo conjunto entre el que proporciona los datos, el operador móvil, y el INE, que es el garante metodológico, el que tiene que aplicar la metodología para obtener las variables que son requeridas, por ejemplo, por los reglamentos europeos. Con lo cual, lo que tenemos es una asociación, es un partenariado entre ambos, que nos obliga a trabajar conjuntamente para resolver esos problemas. Y que eso va cambiando con el tiempo, porque, por ejemplo, te podría decir que hace 10 años, las únicas tarjetas SIM que podíamos detectar en las entradas por la frontera eran las tarjetas SIM de aparatos móviles, de teléfonos, de tabletas, de ordenadores.

Pero, sin embargo, hoy en día cada día es mayor el número de electrodomésticos o de aparatos que disponen de tarjetas SIM, y naturalmente no son turistas. Con lo cual, tenemos que ser capaces de discriminar esa información.

Suelí Lorenzo

Alfredo, en tu opinión, ¿estas fuentes de datos van a sustituir totalmente a las fuentes tradicionales? ¿Dejaremos de preguntarle al ciudadano?

Alfredo Cristóbal

Bueno, es difícil responder de cara al futuro, pero lo que sí que es cierto, y lo que creo que podemos tener todos claro, es que este tipo de fuentes se van a ir utilizando cada vez más en el tiempo.

Pero, como he dicho al principio, hay determinadas variables que no se recogen ni en los registros administrativos, ni en el caso de las fuentes innovativas, como puede ser la telefonía móvil o el *web scaping*. Por ejemplo, te podía citar, con los datos de la telefonía móvil somos capaces de conocer el número de turistas que entra a España en determinado mes. Podemos saber sus etapas, la duración de sus etapas, pero lo que nunca vamos a poder conocer, por ejemplo, es el motivo.

Es decir, qué es lo que le motivó para hacer este viaje. Si es por ocio, si es por negocios, si es por un tema familiar o cualquier otro motivo. Es decir, eso, para poder conocerlo, no nos va a quedar más remedio que seguir preguntándoselo. Por lo tanto, resulta un poco complicado pensar en que en el futuro podríamos prescindir de estas encuestas de tipo tradicional.

Suelí Lorenzo

En una conversación sobre innovación, hoy en día es imposible no referirse a la inteligencia artificial. ¿Qué papel tiene en el INE?

Alfredo Cristóbal

La inteligencia artificial a día de hoy es el gran reto, pero también la gran oportunidad para la producción estadística.

Estamos utilizando procedimientos en el cual somos capaces de gestionar grandes volúmenes de datos, somos capaces de proporcionar unos entornos muy fáciles para la recogida de información automática y, por lo tanto, sabemos que cada día más estamos utilizando procedimientos ligados a la inteligencia artificial. Es fácil hoy en día observar procedimientos de aprendizaje automático, de *machine learning*, en muchísimas encuestas y, por lo tanto, en muchísimas estadísticas y, por lo tanto, esto irá seguramente de más a más en el futuro. Lo que sí que es cierto es que cada día nos sorprende más, es decir, que cada día somos capaces de ver nuevas aplicaciones, lo cual nos proporciona un panorama interesante para el futuro próximo.

Suelí Lorenzo

Alfredo, entiendo que todo esto exigirá perfiles profesionales muy distintos a los que se venían demandando hace unos años.

Alfredo Cristóbal

Efectivamente, Suelí, realmente el INE a día de hoy necesita profesionales capaces de analizar grandes volúmenes de datos. Digamos que necesitamos perfiles asociados a la ciencia de datos.

Eso no quiere decir que también tengamos que conseguir formar a las personas que están trabajando a día de hoy dentro del INE, digamos, para poder conseguir que también puedan ser eficientes dentro de estas técnicas. Lo que sí que es cierto es que detrás de una estadística tiene que haber siempre un buen profesional que sea capaz de analizar la información y de garantizar su calidad.

Suelí Lorenzo

Es decir, que la tecnología ayuda, pero no sustituye la responsabilidad de quienes producen la estadística.

Alfredo Cristóbal

Efectivamente, a través de procedimientos automáticos o de la inteligencia artificial podemos ser más eficientes, eso no cabe duda. Pero lo que es cierto es que hay determinados aspectos, como los ligados a la transparencia, como los ligados a la difusión, como los ligados a la gestión de las normas éticas, que, aunque sea tentador dejarlo en manos de la inteligencia artificial, siempre conviene, por lo menos desde mi punto de vista, que haya una persona detrás de ello.

Suelí Lorenzo

La innovación no sólo cambia la manera en cómo se obtienen y se procesan los datos, también permite responder preguntas que antes eran difíciles de contestar. Un ejemplo son las estadísticas experimentales. ¿Podrías, Alfredo, contarnos un poco en qué consisten?

Alfredo Cristóbal

Bueno, las estadísticas experimentales nacen justamente a la vez que empezamos a recoger información de forma automática, de una manera de la cual no estábamos acostumbrados a gestionar y, por lo tanto, lo que necesitábamos era experimentar hasta poder extraer toda la información disponible que nos permitiera ser capaces de producir una estadística. Eso es lo que dio lugar a las estadísticas experimentales.

Suelí Lorenzo

¿Y puede una estadística experimental acabar siendo una estadística oficial?

Alfredo Cristóbal

Teóricamente, ese tiene que ser su destino real. Puesto que un fenómeno que se comienza a medir, que continúa siendo relevante y que hemos trabajado para poder

garantizar su calidad, lo más lógico es que termine siendo o formando parte de la base estadística del INE. Pero bueno, también puede ser que el fenómeno que se quiso medir en su momento pierda su relevancia en el tiempo, en cuyo caso podríamos dejar de publicarla, podríamos dejar de realizarla sin ningún tipo de problema. Es una estadística experimental.

Suelí Lorenzo

Acabamos de comentar que la innovación permite también responder a nuevas preguntas. ¿Qué fenómenos actuales crees que necesitan todavía de nuevas formas de medición?

Alfredo Cristóbal

Bueno, yo creo que hoy en día, y seguramente también llevándonos al futuro próximo, ámbitos como el medio rural y el medio ambiente son susceptibles de poder mejorar muchísimo su tratamiento a través de este tipo de fuentes.

Podemos ver cómo, por ejemplo, en el medio rural cada vez hay más dispositivos, cada vez más sensores que nos pueden medir, ya no sólo la temperatura, sino también el estado de maduración de un cultivo o simplemente la disposición de una cabaña ganadera a través de un pasto y, por lo tanto, poder también ser susceptibles de poder utilizarse en el ámbito estadístico. En el medio ambiente, pues lo mismo, tenemos cada vez más dispositivos que nos miden en tiempo real y de forma continua la calidad del aire, la contaminación o las condiciones atmosféricas. Realmente tenemos ahí ámbitos en los cuales estoy convencido de que, en un futuro próximo, y ya en la actualidad desde luego, tenemos que empezar a trabajar y que ahí van a tener muchísimo impacto este tipo de fuentes.

Suelí Lorenzo

Alfredo, una última cuestión relacionada con la confianza. ¿Puede la innovación ayudar a que la sociedad entienda mejor cómo se producen las estadísticas oficiales?

Alfredo Cristóbal

Estamos en una época de transición y, por lo tanto, estamos cada vez utilizando más este tipo de técnicas, lo cual ayuda a que la sociedad nos vea cada vez más como una institución altamente tecnológica y cualificada, que nos permita de forma muy rápida responder a necesidades emergentes. Lógicamente, necesitaremos continuar desarrollando metodologías solventes que nos permitan garantizar la calidad de lo que estamos produciendo.

Suelí Lorenzo

Alfredo, hoy hemos visto que detrás de cada dato hay mucho más que un número. Hay métodos que evolucionan, hay nuevas fuentes de información y profesionales que trabajan para medir una realidad cada vez más compleja. Para terminar, y mirando al futuro, si pensamos en el INE dentro de 10 años, ¿qué crees que habrá cambiado en la forma de producir estadísticas?

Alfredo Cristóbal

Buena pregunta. Realmente, en un futuro próximo, yo lo que veo es que estos procedimientos innovativos se van a ir utilizando cada vez más, pero que seguirán teniendo su hueco los registros administrativos y también las encuestas para cubrir aquella información que no podemos obtener de otra manera. Seguramente, el reto no será el utilizar cada vez más un determinado tipo de métodos u otros, sino su integración. Es decir, para nosotros la clave del éxito va a ser el saber cómo integrar los tres tipos de fuentes, tradicionales, registros administrativos, fuentes privadas, en una operación estadística que ofrezca la mejor información para el ciudadano, la información más útil con la que sea capaz de tomar decisiones.

Suelí Lorenzo

Alfredo Cristóbal, director general de Estadísticas Económicas del INE, muchas gracias por acompañarnos.

Alfredo Cristóbal

Muchas gracias a ti, Suelí.

Suelí Lorenzo

Hasta aquí este episodio de En Clave INE.

Puedes acceder a todos nuestros podcasts en www.ine.es así como en YouTube, Spotify o en iVoox. También puedes encontrar en nuestra página web información adicional sobre los temas tratados hoy. Gracias por escucharnos.