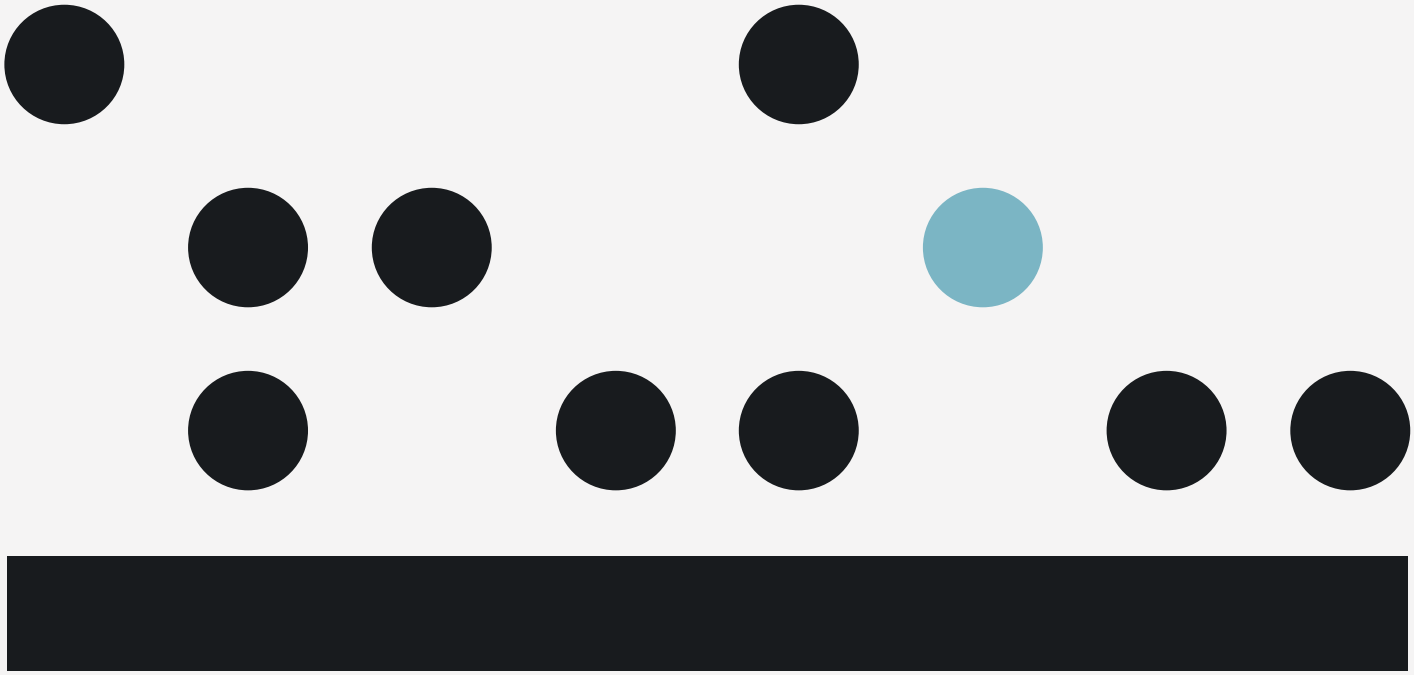
De Ushuaia a La Quiaca: byte por byte

Cómo potenciar el sector turístico con big data

Datos



Daniel Yankelevich
Elián Soutullo
Juan Gabriel Juara
Juan Manuel Ortiz de Zárate
Juan Pablo Ruiz Nicolini
Mariana Kunst
Mariano Zapatero
Rafael Grimson
Rodrigo Castro

Julio 2023

Ministerio de Turismo y Deportes Argentina

SEDLab Simulación de Eventos Discretos .UBAEXACTAS

fundar

De Usuhaia a La Quiaca: byte por byte

Cómo potenciar el sector turístico con big data

Daniel Yankelevich
Elián Soutullo
Juan Gabriel Juara
Juan Manuel Ortiz de Zárate
Juan Pablo Ruiz Nicolini
Mariana Kunst
Mariano Zapatero
Rafael Grimson
Rodrigo Castro

- Generar riqueza
- Promover el bienestar
- Transformar el Estado

Índice

De Usuahia a La Quiaca: byte por byte

Cómo potenciar el sector turístico con big data

fundar

Introducción

¿Cómo generar una herramienta para conocer el público objetivo de las políticas de un ministerio usando big data? ¿Puede una fuente alternativa de datos como esta coexistir con las fuentes tradicionales y complementar la información con la que hoy se cuenta? El presente documento intenta responder a esas preguntas presentando una experiencia colaborativa entre el Área de Datos de Fundar, la Dirección Nacional de Mercados y Estadística Ministerio de Turismo y Deportes de la Nación y el Laboratorio de Simulación de Eventos Discretos de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. En este trabajo conjunto se buscó, a partir del uso de big data, profundizar en el análisis y entendimiento de los flujos del turismo interno y generar nueva información en torno a mercados turísticos para el desarrollo de políticas públicas informadas y basadas en evidencia. Se intentó demostrar, además, cómo los datos generados a través de herramientas de big data se pueden combinar con la información administrativa y de encuestas disponible para enriquecer el diseño y la implementación de esas políticas.

En efecto, el Estado necesita contar con información para la toma de decisiones, el diseño de políticas públicas efectivas y la evaluación de su impacto. En particular, se requieren datos sobre la población a la que se dirigen esas políticas y el contexto en el que se implementarán. Sin embargo, el uso de big data y de fuentes alternativas de información no es una práctica frecuente en la administración pública, que por lo general restringe su uso a "recortes" de información.

Sabemos que los datos con los que se cuenta en el ámbito estatal usualmente son imperfectos e incompletos. A modo de ejemplo, al momento de planificar una política pública tan importante y exitosa como fue el Ingreso Familiar de Emergencia (IFE) en 2020, durante la pandemia de COVID-19, se estimó que alcanzaría aproximadamente a 4 millones de hogares, pero al momento de implementarla se anotaron 13,4 millones (luego del análisis de requisitos, quedaron 8,9 millones)[1].

Para realizar estimaciones con mayor precisión, los datos de fuentes más tradicionales podrían complementarse con fuentes alternativas que permitan tomar decisiones sobre la base de más parámetros y de manera más ajustada al contexto. En muchos casos estas fuentes alternativas se descartan por poco confiables o porque se vuelve difícil "insertarlas" en el análisis, ya que se deben ensamblar de manera coherente con otros datos.

En el caso del Ministerio de Turismo y Deportes se trabaja en el desarrollo de una herramienta informática que tiene como objetivo recopilar información que se genera a nivel subnacional, como complemento del relevamiento que el organismo realiza a partir de un operativo estadístico (Encuesta de Viajes y Turismo de los Hogares - EVyTH). El proyecto avanza con la incorporación de tecnología y la modernización de la plataforma de recopilación, filtrado y análisis, para armonizar los datos que se generan y publican en las distintas jurisdicciones. Sin embargo, dado que estas dos fuentes no cubren todas las necesidades de información, se pensó en complementarlas con los datos provistos por otras fuentes, que pueden proporcionar un punto de vista alternativo, por ejemplo, en cuanto al nivel de desagregación territorial y temporal. Agregar una nueva fuente a los datos que se utilizan implica el desafío de entender la información en contexto, evaluar su calidad, "limpiarla" y, sobre todo, entender en qué aspecto complementa a la información existente.

Presentación del problema: big data para el turismo

En este trabajo se presenta una experiencia práctica para desarrollar una metodología que incorpora fuentes de datos alternativas como complemento a fuentes tradicionales, tales como encuestas y registros administrativos. Como describiremos, la fuente alternativa usada es una base de datos georreferenciada que se recopila a partir de dispositivos móviles y brinda una oportunidad para explorar

1 D'Alessandro, Mercedes (2022). Ingreso Familiar de emergencia: una política pública a contrarreloj. Buenos Aires: Fundar.

datos que los operativos estadísticos no contemplan: aporta información de origen de los viajes a nivel de radios censales para todo el territorio nacional, con granularidad temporal diaria y referencia desagregada sobre los destinos de esos viajes. Si bien sabemos que la información contiene sesgos y ruido, el nivel de granularidad que aporta es significativo. Partes de las conclusiones de este trabajo son específicas para esta aplicación y algunas de las técnicas utilizadas tienen foco en esta base, pero por lo general la metodología puede extrapolarse a otras fuentes y contextos de aplicación.

En la primera sección del documento se presentan las características de los datos, la infraestructura necesaria para el procesamiento y los algoritmos. En la segunda sección se compara el comportamiento turístico a partir de los datos de big data con información de encuestas y registros administrativos. En la tercera sección se presentan casos de uso de la herramienta: para fiestas, turismo de naturaleza y exploración de perfiles de turistas a partir del análisis de aplicaciones de celular. Por último, en la cuarta sección se presenta un resumen de la metodología implementada.

Presentación del problema: big data para el turismo

Datos, infraestructura y algoritmos

El punto de partida de la experiencia que aquí relatamos incluyó un trabajo en conjunto con el equipo de datos del Instituto Nacional de Promoción Turística (INPROTUR), que puso a disposición una base de datos y su experiencia utilizándolos en otros proyectos. De esa manera, se elaboraron algunas hipótesis preliminares que allanaron el camino para el posterior análisis realizado en este proyecto. Esto es de especial valor cuando la documentación provista por la fuente original de los datos resulta insuficiente para cubrir todas las interpretaciones de la información que requiere el proyecto.

La base de datos de Inprotur a partir de la que se trabajó es una fuente especializada para el trabajo de *marketing* digital, que contiene información anónima de dispositivos móviles identificados a partir de un código único destinado a publicidad o IFA (*Identifier for Advertising)*[2]. Es decir, que no hay posibilidad de asociar registros de datos con los individuos a los que esos registros se refieren. Por lo tanto, la base de datos no contiene datos personales ni se pueden reidentificar porque no hay elementos que asocien un registro a una persona individual.

> Los documentos de trabajo Anónimos pero no tanto: cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad y La anonimización: un instrumento clave para una gestión de datos eficiente profundizan sobre el problema de compartir información sin violar el derecho a la privacidad de las personas. Todo dato es sensible si causa un daño a la persona en el momento en que se hace público. Los datos pueden ser considerados sensibles por muchos motivos: origen racial y étnico, opiniones políticas, patrimonio personal, convicciones religiosas, afiliación sindical o información referente a salud o vida sexual.

La base de datos sobre la que se trabajó está conformada por tres conjuntos de datos principales (GEO, CEL-CDL y APPS) con un total de más de 6700 millones de registros (renglones de información) almacenados en un volumen de 300 GB (aproximadamente el tamaño de 100 películas).

---

2 Para mayor detalle puede consultarse: Guidelines For Identifier For Advertising On OTT Platforms.

Presentación del problema: big data para el turismo

La información refiere a aproximadamente 17 millones de dispositivos únicos (equivalente —en promedio— a 1 dispositivo cada 2 personas mayores de 15 años en la Argentina) con registros georreferenciados diariamente, durante un período de 12 meses continuos (los previos al inicio del aislamiento social preventivo y obligatorio —ASPO— establecido por la llegada del COVID-19 a la Argentina, es decir desde abril de 2019 hasta marzo de 2020).

> La información refiere a aproximadamente 17 millones de dispositivos únicos (equivalente en promedio a 1 dispositivo cada 2 personas mayores de 15 años en la Argentina) con registros georreferenciados diariamente, durante un período de 12 meses continuos, desde abril de 2019 hasta marzo de 2020; la cobertura geográfica alcanza a todas las provincias.

La cobertura geográfica de los datos alcanza a todas las provincias, incluyendo personas cuya residencia puede asociarse a 528 departamentos (todos los de nuestro país menos la Antártida e islas del Atlántico Sur) y que han realizado viajes que abarcan los mismos 528 departamentos mencionados (con un total de más de 41 millones de identificaciones de movilidad diaria hacia departamentos distantes del departamento de residencia de las personas). No todos estos traslados corresponden a destinos turísticos o destinos en general: se trata de la movilidad de la persona, que podría estar en tránsito hacia un destino más lejano. Es importante señalar la diferencia entre datos que indican movilidad y datos asociados a un destino con fines turísticos.

Estos datos poseen algunas características típicas de big data. En primer lugar, su volumen es lo suficientemente grande como para que los algoritmos que se diseñen para obtener respuestas a partir de los datos puedan devolver resultados en tiempos razonables (desde algunas horas hasta quizá algunas decenas de horas, según el caso) con una infraestructura de cómputo adecuada . En segundo lugar, la generación y agrupamiento original de los datos no fueron diseñados teniendo en mente el uso que se le daría en este trabajo, ni se tuvo control alguno sobre la representatividad estadística de los registros recolectados. En tercer lugar, el ejercicio de recolección de datos no es auditable ni repetible bajo condiciones controladas.

En síntesis, "es lo que hay, es muy grande y hay que determinar qué es lo que se puede —y lo que no se puede— hacer con ello, en un contexto de tiempos y recursos acotados". Podríamos decir: "un día típico en la vida de big data".

Para enfrentar estos desafíos fue necesario trabajar de manera interdisciplinaria, combinando las experiencias de expertos y expertas en el campo del turismo, en análisis estadístico de la información, en diseño de algoritmos eficientes, en política pública y en sociología, por mencionar algunas disciplinas claves.

Dado el gran volumen de la información ya mencionado, las "preguntas interesantes y relevantes" que pueden realizarse a los datos —esperando respuestas en un tiempo acotado, como se dijo— deben ser consideradas con cuidado, ya que del tipo de pregunta dependerá la adecuación de los algoritmos y el tiempo que tomará su procesamiento en infraestructuras de cómputo especiales. Por ello, un proyecto de big data del cual se esperan respuestas con potencial para ser utilizadas como insumo de políticas públicas exige un equilibrio interdisciplinario particular.

La estructura del proyecto requirió organizar equipos de personas trabajando distribuidas geográficamente, de manera simultánea e independiente pero coordinada, procesando distintas porciones de los datos.

Presentación del problema: big data para el turismo

Mientras un equipo podía estar redefiniendo o adecuando filtros que fueran adecuados para, por ejemplo, definir los límites geográficos para contabilizar visitas a un Parque Nacional, otro equipo podía estar ajustando los métodos de análisis del perfil socioeconómico de los visitantes a dicho parque en base a sus departamentos de origen. Luego, los resultados de los equipos podían cruzarse, lo que hacía necesario reprocesar datos. Este esquema de trabajo es viable, pero requiere de una metodología consistente. Para ello definimos una jerarquía acorde a las distintas necesidades de potencia de procesamiento de datos: nivel big data (servidores de muy alta capacidad, procesamiento de datos originales, con cada señal individual producida por cada dispositivo durante los 12 meses), nivel intermedio (servidores de capacidad intermedia, datos reducidos —por ejemplo a granularidad diaria y por departamento—) y nivel de análisis (PC de oficina, datos reducidos para análisis puntuales —por ejemplo, datos relacionados sólo a visitas a fiestas nacionales—).

Este esquema tipo árbol permitió a los equipos trabajar en paralelo y a diferentes ritmos, e incluso adaptarse de manera flexible según la exigencia de cada etapa de avance del proyecto. Para una coordinación eficiente y eficaz en un contexto distribuido, se usaron herramientas de colaboración remota en tiempo real (para compartir y editar documentos de manera colaborativa, y tener acceso remoto y simultáneo a bases de datos compartidas)

Caracterización, calidad, representatividad y sesgos de los datos

Para analizar la calidad y los sesgos de los datos georreferenciados nos concentramos en aquellos que tienen definida una *"common evening location"* (CEL), que es la ubicación típica durante la noche de ese dispositivo. Se asume como criterio que esa ubicación típica es la residencia de la persona identificada por el dispositivo.

La coordenada geográfica provista por las CEL nos permite asociar cada IFA a un radio censal (unidad administrativa de área geográfica con mayor granularidad según la cual el INDEC organiza el censo nacional cada 10 años). Esto permite asociar el IFA al conjunto de características socioeconómicas provistas por el INDEC para su zona de residencia.

Para este trabajo, contamos en total con unos 17 millones de IFA únicos de partida a lo largo de los 12 meses disponibles. Sin embargo, sólo 12,9 millones de ellos cumplían con requisitos mínimos de calidad para ser considerados útiles en los análisis posteriores (por ejemplo, cada IFA debía tener un CEL bien definido, con una geolocalización razonable dentro del territorio nacional).

Sin embargo, la cantidad de datos no fue pareja ni espacial ni temporalmente. Por ejemplo, hubo meses con datos de casi 7 millones de IFA con CEL bien definido (2019-05), mientras que otros meses (2020-01) apenas alcanzaron los dos millones.

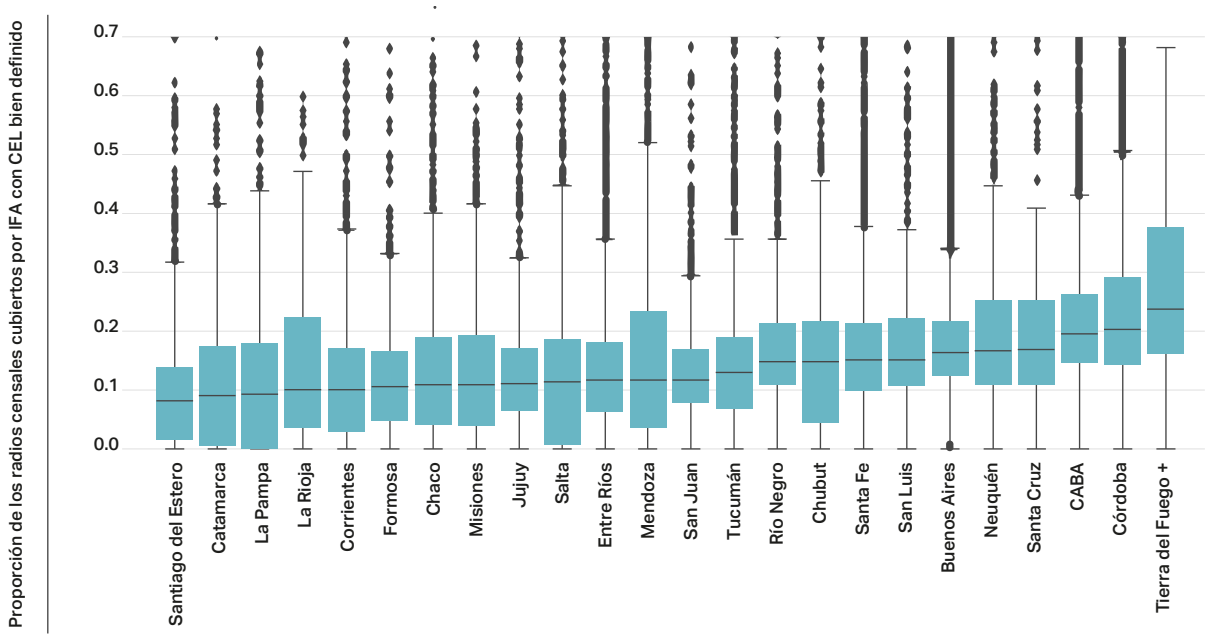
A nivel espacial también se apreciaron diferencias, aunque fueron de menor envergadura que las temporales. El siguiente gráfico muestra, para cada provincia, un gráfico de cajas que resume la distribución de la proporción de los radios censales de cada provincia abarcados por aquellos IFA que tuvieran un CEL bien definido en el conjunto de datos.

Gráfico 1

Presentación del problema: big data para el turismo

Residencia de los usuarios de la muestra de big data. Proporción de los radios censales cubiertos por los dispositivos móviles (IFA) de acuerdo a su ubicación típica durante la noche (CEL bien definido) por provincia

Gráfico 1

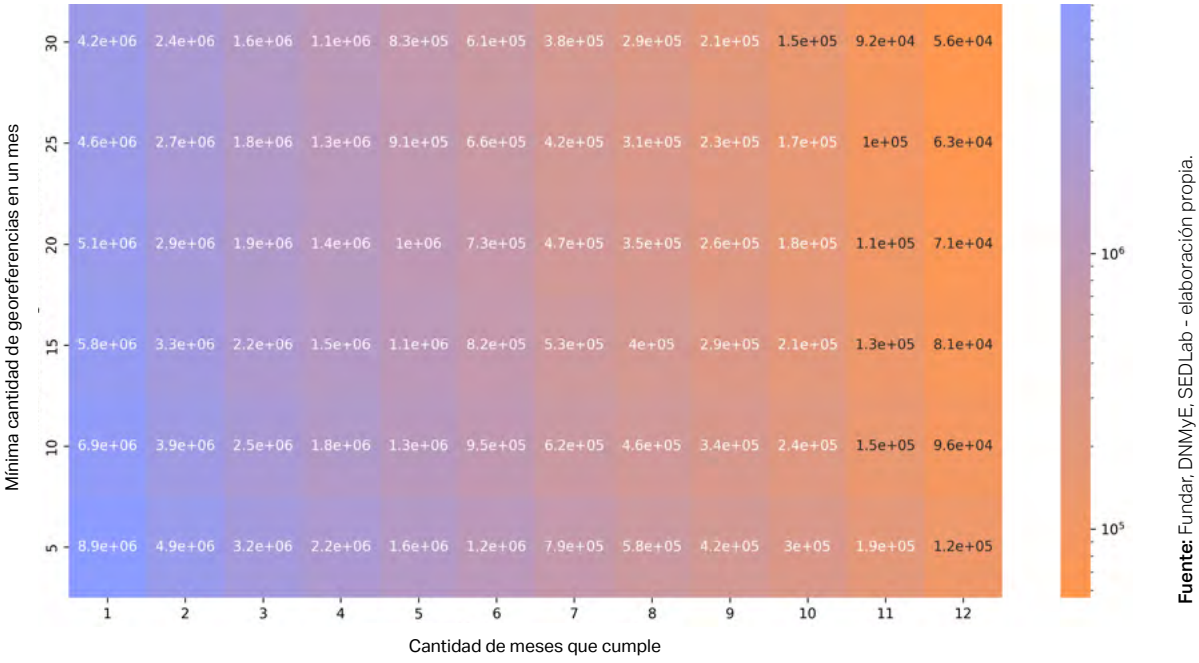


Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Por otra parte, se observa que es usual contar con datos de un IFA solo para algunas fechas mientras que es poco frecuente contar con muchos datos georreferenciados para un mismo IFA a lo largo de todo el año. La siguiente tabla resume la cantidad de IFA que cumplen con la condición de tener varios meses con una cantidad mínima de datos georreferenciados cada mes.

Volumen de georeferencias disponible para los usuarios de la muestra de big data. Número de dispositivos móviles (IFA) que cumplen con la cantidad mensual mínima de georeferencias, por una cantidad dada de meses (1 a 12), durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 2



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Presentación del problema: big data para el turismo

Nuestro objetivo fue buscar IFA que nos brindaran información confiable o de calidad. En el ámbito de los datos, "calidad" no es un concepto absoluto: en este contexto, quiere decir que cuente con varios meses que contengan datos y con suficiente información georreferenciada dentro del mismo mes. Para definir más precisamente qué quiere decir varios meses y suficiente información es necesario establecer límites numéricos. De esta forma se pueden identificar claramente los datos a utilizar en el análisis, y eliminar datos irrelevantes o que introducen ruido.

A medida que se incrementa el requerimiento de calidad mencionado, la cantidad de IFA disponibles disminuye sensiblemente. Si consideramos que un IFA debe satisfacer la condición de georreferenciación un par de veces en al menos un mes, el número es grande, pero si aplicamos una condición más restrictiva, la muestra se reduce de manera considerable.

> Nuestro objetivo fue buscar IFA que nos brindaran información confiable o de calidad, pero en el ámbito de los datos, "calidad" no es un concepto absoluto: en este contexto, quiere decir que cuente con varios meses que contengan datos y con suficiente información georreferenciada dentro del mismo mes.

Hubo 8,9 millones de usuarios que fueron georreferenciados al menos 5 veces en un mes, pero solo unos 120.000 que fueron georreferenciados 5 veces todos los meses. Análogamente, hubo 4,2 millones que fueron georreferenciados al menos 30 veces en un mes, pero solo unos 150.000 que fueron georreferenciados 30 veces al menos 10 meses dentro del año.

En función de estas observaciones decidimos analizar tres muestras de datos, que utilizamos en las verificaciones de sesgo y calidad:

- **Muestra 1**: aquellos IFA que tuvieron al menos 5 georreferenciaciones durante al menos un mes.
- **Muestra 2:** aquellos IFA que tuvieron más de 15 georreferenciaciones durante más de 6 meses.
- **Muestra 3:** aquellos IFA que tuvieron más de 30 georreferenciaciones durante más de 10 meses.

Para analizar posibles sesgos de representación y entender si las conclusiones que se tomaran con estos datos podían generalizarse a toda la población, comparamos estos datos con los del último censo nacional publicado en el momento de hacer el estudio (2010). La comparación se realizó eligiendo algunas métricas definidas a nivel de radio censal: el Índice de Vulnerabilidad Sanitaria (IVS) y el Nivel Socioeconómico (NSE) confeccionados por la fundación Bunge y Born[3].

A continuación presentamos dos gráficos con los resultados de este análisis:

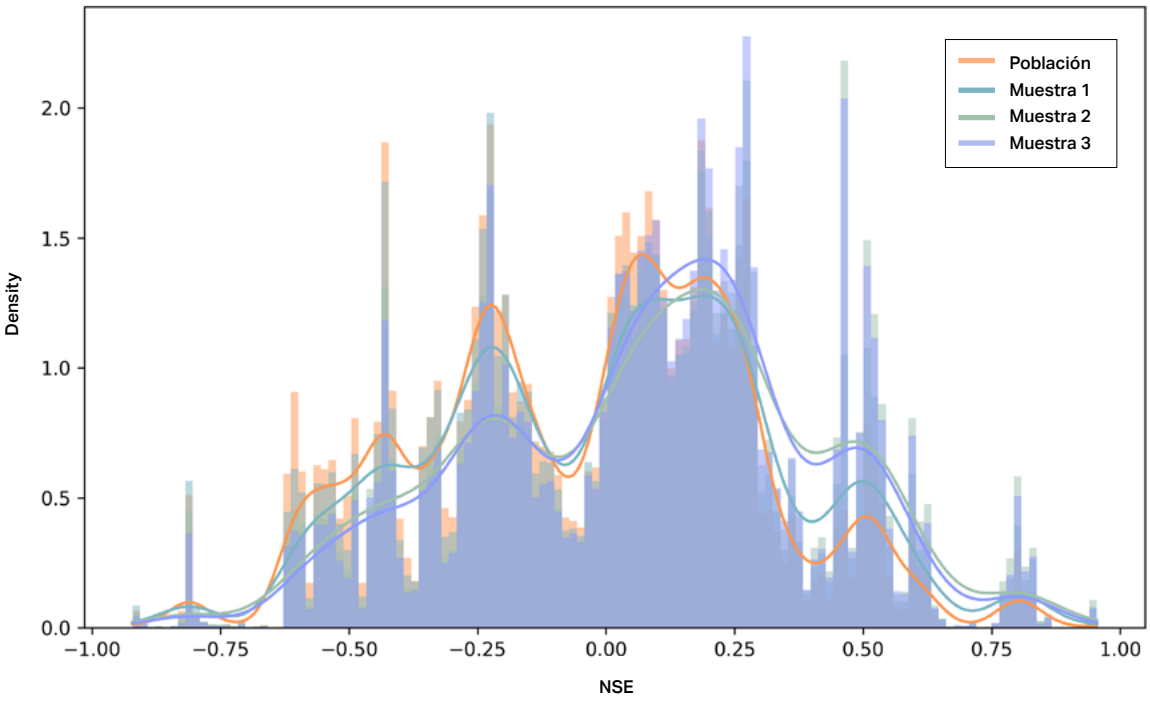
Gráfico 3

3 https://www.fundacionbyb.org/_files/ugd/2aae47_18cd119620bc4b16aae2d643cf416af8.pdf?index=true

Presentación del problema: big data para el turismo

Representatividad de la muestra de big data (NSE). Comparación de tres muestras de big data con la población argentina (INDEC 2010) según el Nivel Socioeconómico (NSE)

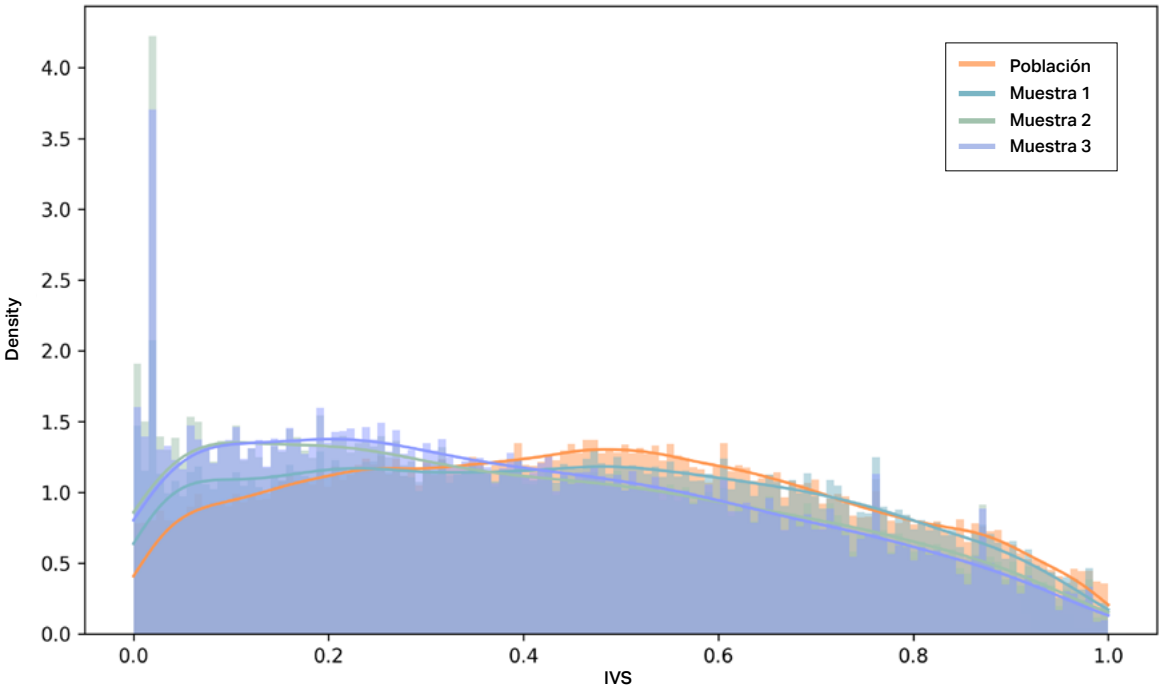
Gráfico 3 a



Fuente: Fundar, DNMyE, SEDLab - elaboración propia a partir de datos proporcionados por el INDEC y la Fundación Bunge Y Born.

Representatividad de la muestra de big data (IVS). Comparación de tres muestras seleccionadas de big data con la población argentina (INDEC 2010) según el Índice de Vulnerabilidad Sanitaria (IVS)

Gráfico 3 b



Fuente: Fundar, DNMyE, SEDLab - elaboración propia a partir de datos proporcionados por el INDEC y la Fundación Bunge Y Born.

Se puede observar que la distribución del nivel socioeconómico de la población y de las muestras son bastante similares, lo que indica que las tres muestras presentan una buena representación del nivel socioeconómico de la población.

Vemos que la muestra 1, más numerosa, tiene un mejor ajuste. Aunque su composición evidencia un menor nivel socioeconómico y una mayor vulnerabilidad sanitaria que la población nacional, estas diferencias son sutiles. Al contrario, las muestras 2 y 3 presentan un mayor nivel socioeconómico y una menor vulnerabilidad sanitaria que la población nacional pero, nuevamente, son diferencias sutiles que no permiten inferir un sesgo relevante en las muestras.

Ejemplo de movilidad de un IFA

La identificación del lugar de residencia de los IFA, así como su geolocalización a lo largo del tiempo y del territorio, permiten conocer en detalle su movilidad.

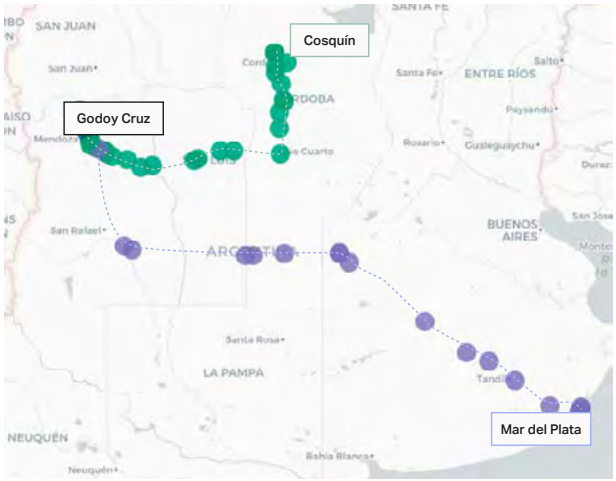
Al analizar datos con estas características, resulta útil complementar el estudio estadístico con una visión cualitativa de casos que permita entender mejor otros aspectos. Por ese motivo, se miraron algunas trayectorias individuales. Por ejemplo, analizamos un IFA con residencia en Godoy Cruz, Mendoza, que tuvo más de 140 registros correspondientes a viajes a más de 20 km de distancia de su hogar. La base de datos nos muestra que esta persona pasó 91 días viajando durante los 12 meses del período de referencia.

En el siguiente mapa se observan dos de los principales viajes que realizó: en el mes de enero visitó Cosquín (Córdoba), en la misma fecha que se realizó el Festival Nacional de Folklore[4]; mientras que entre la segunda y tercera semana de febrero las ubicaciones del IFA muestran claramente un viaje a Mar del Plata.

Este tipo de ejercicio brinda la posibilidad de analizar no solo los destinos potenciales de los IFA, sino también el recorrido realizado para desplazarse hacia ellos, con una perspectiva que complementa la visión estadística puramente cuantitativa.

Gráfico 4

Ejemplo de movilidad de un dispositivo móvil. Registros de geolocalización de un dispositivo (IFA) con residencia en Godoy Cruz (Mendoza), correspondientes a desplazamientos a más de 20 km de distancia de su hogar, durante el período de referencia (abril 2019 a marzo 2020)



Registros de ubicación en enero 2020 (misma semana que el Festival Nacional de Folklore)

Registros de ubicación en febrero 2020 (segunda y tercera semana del mes)

Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

4 Un caso de uso para el que pensamos de utilidad esta herramienta es el de caracterizar visitantes de Turismo de Fiestas. La tercera sección avanza en un ejemplo al respecto.

Viajes y big data

La Dirección Nacional de Mercados y Estadística (DNMyE) del Ministerio de Turismo y Deportes de la Nación es el organismo encargado del Sistema de Estadísticas de Turismo de Argentina[5]. Para ellas, de un modo esquemático, las formas de turismo se clasifican de acuerdo con el origen y destino de los visitantes: interno (quienes se desplazan en el mismo país que residen), receptivo (no residentes que viajan al país) y emisivo (residentes que viajan a otro país).

Clasificación de las formas de turismo de acuerdo con el origen y destino de los visitantes

Tabla 1

Residencia del visitante	Destino del viaje	
	Argentina	Resto del mundo
Argentina	**Interno**	**Emisivo**
Resto del mundo	**Receptivo**	

Fuente: Fundar, DNMyE, SEDLab - elaboración propia, de acuerdo con la clasificación de la Dirección Nacional de Mercados y Estadística (DNMyE) del Ministerio de Turismo y Deportes de la Nación.

Para este trabajo fue de especial interés la información relativa al turismo interno que, en buena medida, es caracterizado a partir de la Encuesta de Viajes y Turismo de los Hogares (EVyTH)[6].

> Desde la DNMyE (Dirección Nacional de Estadísticas y Mercado) se administran diversos operativos estadísticos que tienen como objetivo la caracterización las diferentes formas de turismo: la Encuesta de Turismo Internacional (ETI) y la Encuesta de Ocupación Hotelera (EOH), en conjunto con el INDEC, y la ya mencionada Encuesta de Viajes y Turismo de los Hogares (EVyTH).

A partir de la EVyTH se obtiene información desagregada del comportamiento turístico de los hogares de la Argentina. Con dicha información se pueden analizar las características de la actividad turística de los residentes en nuestro país y, por lo tanto, diseñar e implementar políticas públicas sectoriales con mayor eficacia y eficiencia. Entre sus objetivos principales se destacan la medición de la evolución de los viajes realizados por los hogares o los lugares visitados (además de otras características como formas de alojamiento, gasto o medios de transporte utilizados).

La EVyTH es una encuesta por muestreo, cuya captación se lleva adelante bajo la metodología CATI (*Computer Assisted Telephone Interviewing*). El universo bajo estudio de la encuesta son los grandes aglomerados urbanos definidos en la Encuesta Permanente de Hogares (EPH) del INDEC. Por lo

---

5 Entre los objetivos del Sistema de Estadísticas de Turismo está , por ejemplo, el trabajo junto con la Dirección de Cuentas Nacionales del Instituto de Estadísticas y Censos (INDEC) para la publicación de la Cuenta Satélite de Turismo de Argentina (CST -A) a partir de los lineamientos de la Organización Mundial de Turismo (OMT).

6 El operativo estadístico de la EVyTH se realiza de manera continua desde el año 2012 con publicaciones trimestrales. Puede consultarse acá el Documento Metodológico y acceder a los microdatos de la encuesta en el Portal de Datos Abiertos del Sistema de Información Turística de la Argentina (SINTA).

tanto, se consideran 32 aglomerados (o GAU, por Grandes Aglomerados Urbanos) agrupados en 8 regiones (Ciudad Autónoma de Buenos Aires, partidos del Gran Buenos Aires, resto de la Provincia de Buenos Aires, Córdoba, Litoral, Norte, Cuyo y Patagónica). El tamaño de muestra mensual es de 2600 hogares distribuidos entre todos los aglomerados urbanos. La ventana de observación es bimensual, esto es, se indaga sobre los viajes realizados durante los últimos dos meses. De esta forma, con el mecanismo de períodos de referencia y ventanas de observación implementados se logra que, para un mes determinado, se disponga de 5200 casos para realizar las estimaciones de ese mes.

Superar limitaciones metodológicas con fuentes de datos no convencionales

El uso de fuentes alternativas de datos, a partir de un método como el propuesto en este documento de trabajo, nos da la posibilidad de explorar alternativas para sortear algunas de las limitaciones que el actual diseño metodológico del operativo impone como restricciones. A saber: la EVyTH, por diseño, presenta información estadística agregada a nivel de regiones turísticas, con un recorte temporal trimestral y está pensada para caracterizar el origen (hogares) de los visitantes de los GAU (un 63% de la población argentina, aproximadamente, excluyendo a los residentes de ciudades medianas y pequeñas).

> La fuente de datos georreferenciada a partir de dispositivos móviles nos brinda una oportunidad para explorar información que el operativo estadístico no contempla: información de origen al nivel de departamentos o radios censales para todo el territorio nacional,, con granularidad temporal diaria y referencia desagregada sobre los destinos de esos viajes.

Por el contrario, la fuente de datos georreferenciada a partir de dispositivos móviles nos brinda una oportunidad para explorar información que el operativo estadístico no contempla: información de origen al nivel de departamentos o radios censales para todo el territorio nacional (más allá de los GAU), con granularidad temporal diaria (en vez de trimestral según diseño del operativo, o mensual al explotar los microdatos de las distintas olas) y referencia desagregada sobre los destinos de esos viajes.

En este trabajo, del total de 17 millones registros únicos se pudo asignar residencia a 12,9 millones, de los cuales se identificó que 2,8 millones realizaron viajes. La fuente de datos permitió además determinar que el 59% de estos viajeros reside en GAU mientras que el 41% restante vive en otros lugares (de los que no se tiene información con EVyTH). A su vez, se cuenta con más de 41 millones de registros correspondientes a movimientos diarios de estos viajeros para todo el período de referencia.

¿Qué observamos cuando analizamos el comportamiento turístico a partir de ambas fuentes?

Tal como se explicó en la primera sección, la nueva fuente de información con la que trabajamos permitió identificar usuarios únicos o IFA sobre los cuales se pudo aproximar su lugar de residencia habitual (uno de los requisitos para el análisis del turismo, tal como se realiza en la EVyTH) a partir del lugar más común georreferenciado durante horario nocturno (*CEL*). Esta identificación nos dio la posibilidad de estimar viajes en un sentido análogo al que se estiman para el turismo interno con la EVyTH: se obtiene un primer criterio de validación de un viaje turístico[7] toda vez que un IFA se encuentra a determinada distancia de su residencia

---

7 Siguiendo la definición de *Viaje Turístico* de la EVyTH, al criterio de distancia se le agrega uno de frecuencia de visitas, siendo una repetición semanal el límite para considerar un destino (por cercano o lejano que esté fuera) como parte del entorno habitual. Por otro lado, los viajes a segundas viviendas son considerados por definición como viajes turísticos, por lo que no rige el criterio de habitualidad. Por último, la encuesta excluye desplazamientos en los que los motivos de viajes refieren a actividades remuneradas. Más detalle disponible en el Documento Metodológico - Unidades de observación.

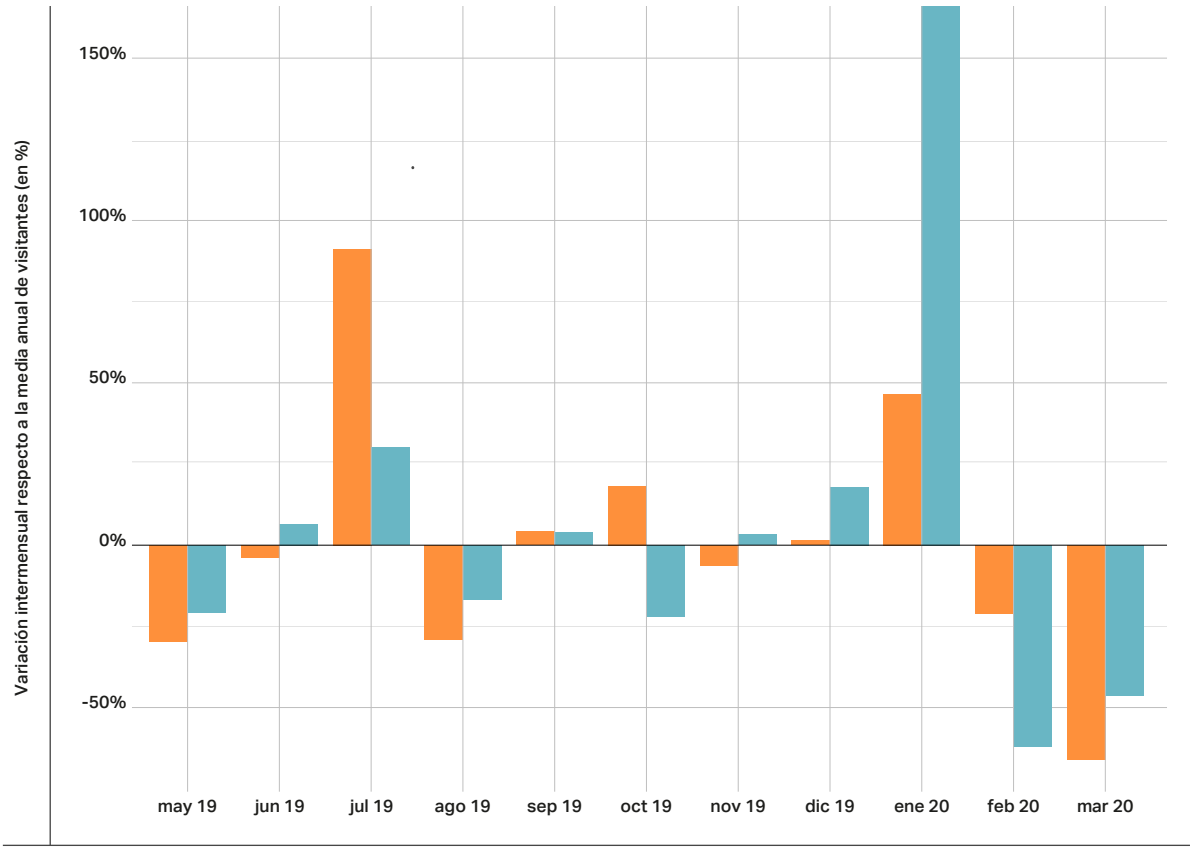
(40 km para el Área Metropolitana de Buenos Aires y 20 km para el resto del país) o entorno habitual.

Resulta oportuno destacar que, como explicamos en el apartado sobre análisis de sesgos de la primera sección, la variación espacial no es tan importante como la variación temporal en la cantidad de datos. Esta última es realmente relevante y requiere ser considerada para poder realizar análisis sobre la variación temporal en la cantidad de visitas a un sitio específico. Por este motivo, definimos un factor multiplicador para la cantidad de visitas registradas cada mes a un destino turístico[8].

Un primer ejercicio consiste en replicar, en la medida de lo posible, el análisis de viajes con la nueva fuente de datos, como si fuera el diseño de la EVyTH. A saber, filtrar datos que corresponden a viajes, del mismo periodo de referencia (abril 2019-marzo 2020), cuyo origen sea alguno de los grandes aglomerados urbanos y su destino cumpla requisitos específicos de distancias recorridas y entorno habitual de los viajeros[9]. El gráfico 5 (que sigue a continuación) muestra la evolución de ambas fuentes de datos que presentan una correlación de 0.73, más allá de los límites ya señalados.

Gráfico 5

Volumen de turismo interno. Evolución de la variación intermensual de visitantes según registros de la Encuesta de Viajes y Turismo de los Hogares (EVyTH) y de big data (IFA), durante los 11 meses del período de referencia (mayo 2019 a marzo 2020)



Fuente: Fundar, DNMyE, SEDLab - elaboración propia, de acuerdo con datos propios e información de la Encuesta de Viajes y Turismo de los Hogares (EVyTH), de la Dirección Nacional de Mercados y Estadística (DNMyE) del Ministerio de Turismo y Deportes de la Nación.

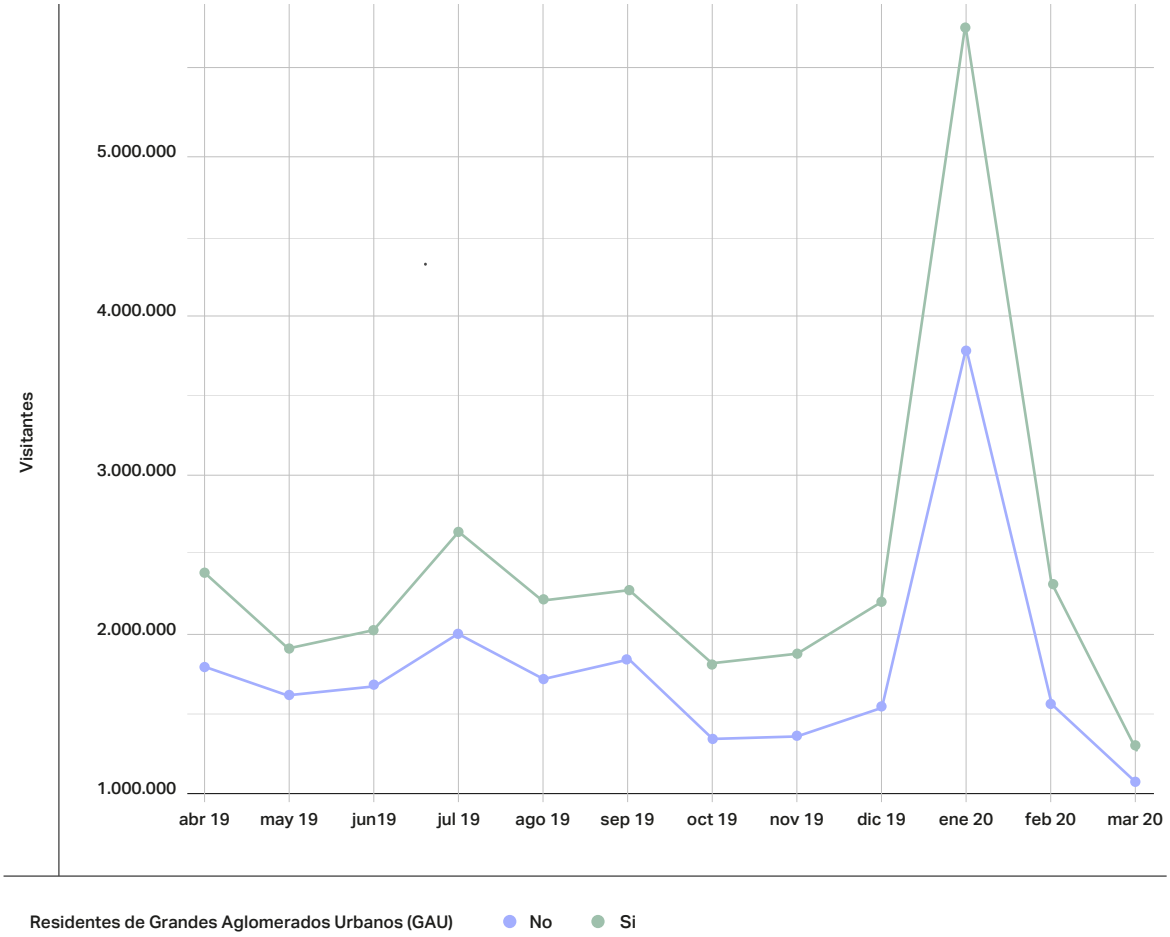
8 El factor, para cada mes, surgió de dividir la población total del CENSO 2010 por la cantidad de IFA únicos con datos georreferenciados dentro del mismo mes. De esta forma, si un mes se registró la mitad de datos que en otro, el factor multiplicador para la cantidad de visitas consideradas como reales será el doble para ese mes.

9 Mientras que en este ejercicio se puede satisfacer el primero de los criterios de la EVyTH (distancia), existen algunas limitaciones para abordar completamente el segundo (entorno habitual) por la falta de información de señales para cada IFA a lo largo del tiempo. No obstante, se trabajó en limitar lo máximo posible el conteo de aquellos viajes de un mismo IFA que se repitieron consecutivamente a lo largo de cuatro semanas al mismo destino. Por último, dada la naturaleza de la información, tampoco tenemos posibilidades de lidiar con el tercero de los criterios que se utilizan en la EVyTH: el filtro de viajes relacionados con actividades remuneradas o trabajo como parte de los traslados, que se desprende de respuestas de los encuestados. Por todo esto, ambas fuentes no son estrictamente comparables y deben ser tomadas como una aproximación.

Cuando extendemos el análisis comparativo al incluir la totalidad de los IFA disponibles y no solo la muestra de los que tienen origen en los GAU se observa, en primer lugar, un comportamiento análogo entre aquellos IFA que residen en GAU y aquellos que no. De estos últimos, la base de datos nos brinda información con la que no se cuenta mediante la EVyTH.

Gráfico 6

Residencia de los visitantes argentinos según big data. Evolución de la cantidad total de visitantes internos (IFA) según si residen (o no) en un Gran Aglomerado Urbano (GAU), durante el período de referencia (abril 2019 a marzo 2020)



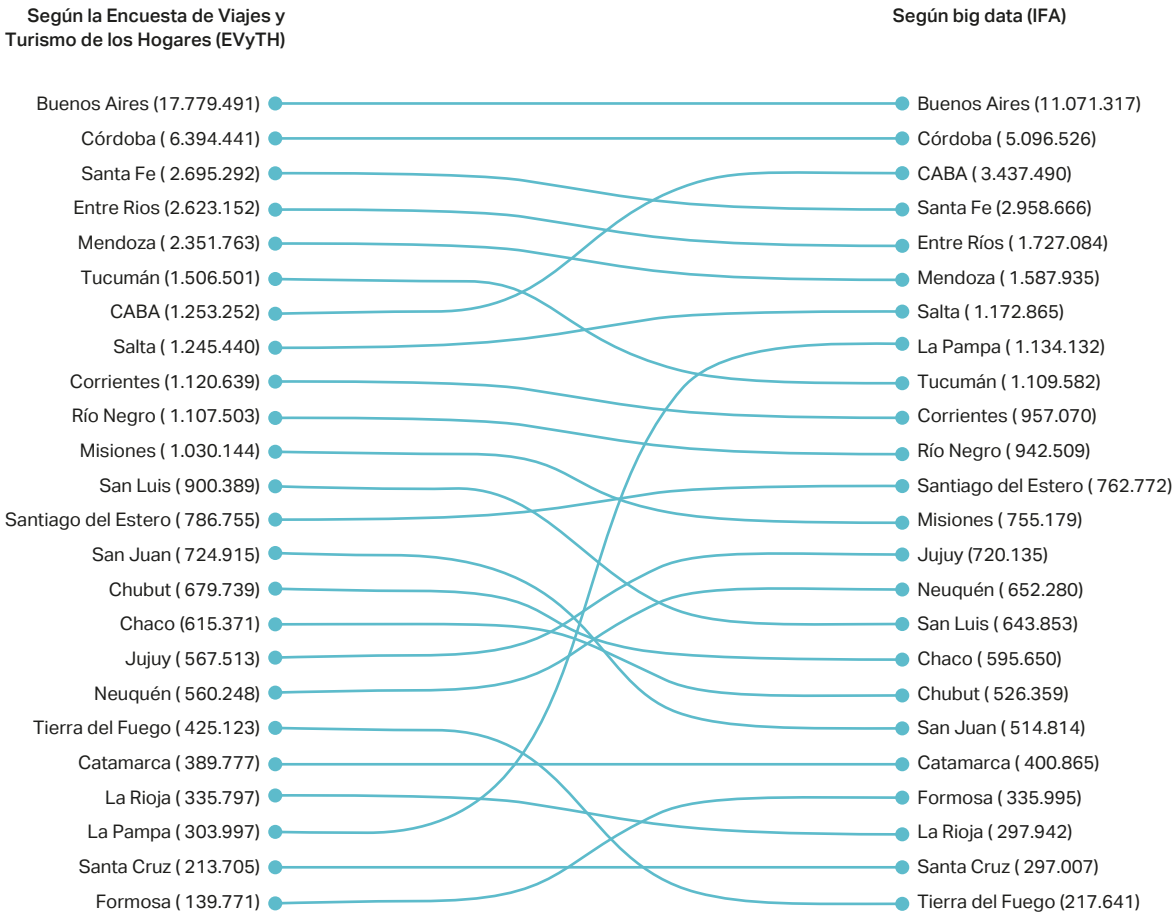
Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Por otra parte, si bien el que podemos estimar es parecido, existen algunas diferencias relevantes para distritos como CABA, La Pampa y Río Negro. En términos estrictos esta divergencia que se observa, *a priori*, podría en realidad estar describiendo fielmente la realidad, en tanto la EVyTH interroga sobre el destino principal de un viaje y los IFA registran una posición geográfica en un determinado momento del tiempo. Que la posición en el ranking de La Pampa sea mucho más alta cuando se trata de registros big data que cuando se observan las respuestas de la encuesta es consistente con que se trate de un "destino de paso" en viajes que tienen como destino final localidades de la Patagonia, por ejemplo[10].

10 Al explorar los datos cualitativamente se observa como, por ejemplo, el departamento Caleu Caleu de La Pampa rankea alto con registros de viajes por IFA (sin ser un destino típicamente turístico) y es a través de este (por la Ruta Nacional 22) una de las principales vías de acceso a la Patagonia desde Buenos Aires.

Ranking provincial de destinos turísticos. Comparación de las provincias argentinas más visitadas según la Encuesta de Viajes y Turismo de los Hogares (EVyTH) y según big data (IFA), de acuerdo a la cantidad total de visitantes internos durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 7



Fuente: Fundar, DNMyE, SEDLab - elaboración propia, de acuerdo con datos propios e información de la Encuesta de Viajes y Turismo de los Hogares (EVyTH), de la Dirección Nacional de Mercados y Estadística (DNMyE) del Ministerio de Turismo y Deportes de la Nación.

Casos de uso de la herramienta

Turismo de naturaleza: visitas a Áreas Protegidas

El turismo de naturaleza o turismo activo natural —que incluye al turismo aventura, el ecoturismo y otras prácticas en las que la naturaleza ocupa un rol protagónico— es un segmento estratégico para el sector turístico de la Argentina y uno de los de mayor crecimiento en nuestro país y el mundo. Este crecimiento refleja una tendencia global en la que los destinos naturales, los espacios abiertos y las experiencias genuinas en ámbitos silvestres permiten a una población crecientemente urbana reconectar con la naturaleza y ocupan un lugar central en la preferencia de los viajeros. La pandemia de COVID-19 no hizo más que reafirmar esta tendencia.

Bajo este marco, desde la Subsecretaría de Desarrollo Estratégico del Ministerio de Turismo y Deportes se desarrolla un programa integral llamado "La Ruta Natural", con el cual la Argentina se propone

Casos de uso de la herramienta

promocionar territorios y comunidades en todo el país para ser receptoras de este tipo de turismo[11].

Una fuente de datos relevante que sirve para hacer seguimiento del comportamiento turístico relacionado con el turismo de naturaleza surge del Registro Nacional de Autorizaciones, Recaudaciones e Infracciones (RENARI) de la Administración de Parques Nacionales (APN)[12], que recopila la cantidad de visitas que recibe cada parque y la condición de residencia de los visitantes (residente o no residente en el país).

Como vimos en la **primera sección,** la determinación aproximada del lugar de residencia de los IFA y el correspondiente nivel socioeconómico que se deriva de datos censales permiten construir perfiles de visitantes sobre los cuales se tiene muy poca información (más allá del volumen de visitas, condición de residencia o categorías de acceso).

El análisis consistió en filtrar del conjunto de datos georeferenciados (coordenadas de latitud y longitud) aquellas señales de los IFA que se encontraban presentes dentro del polígono de áreas protegidas para un periodo de referencia determinado (agrupando las observaciones a nivel mensual, por ejemplo, para comparar con los registros administrativos).

Luego de validaciones con expertos y una serie de procesos de limpieza, el ejercicio evidenció que, si bien el volumen de IFA que visitaron PN es bastante menor al conocido por registros administrativos, presenta un comportamiento similar y provee información hoy no disponible por vía administrativa. Dos cuestiones se destacan: (a) por un lado, por un lado, cuando se aplica un factor normalizador, la cantidad de visitas estimadas en el período, que era de 1% según la información administrativa, pasa a 9,7%; (b) en segundo lugar, se realizaron procesos de limpieza para eliminar "ruido" en las estimaciones, como la posible contabilización de IFA que atraviesan un polígono[13] destacado como "área protegida" pero que son en realidad tránsito de una vía nacional, por ejemplo, y no una visita genuina.

Gráfico 8

11 El Documento de Trabajo N°2 de la Dirección Nacional de Mercados y Estadísticas caracteriza el perfil del Turismo de Naturaleza a partir de la EVyTH (para turismo interno) y la Encuesta de Turismo Internacional (ETI) para el receptivo.

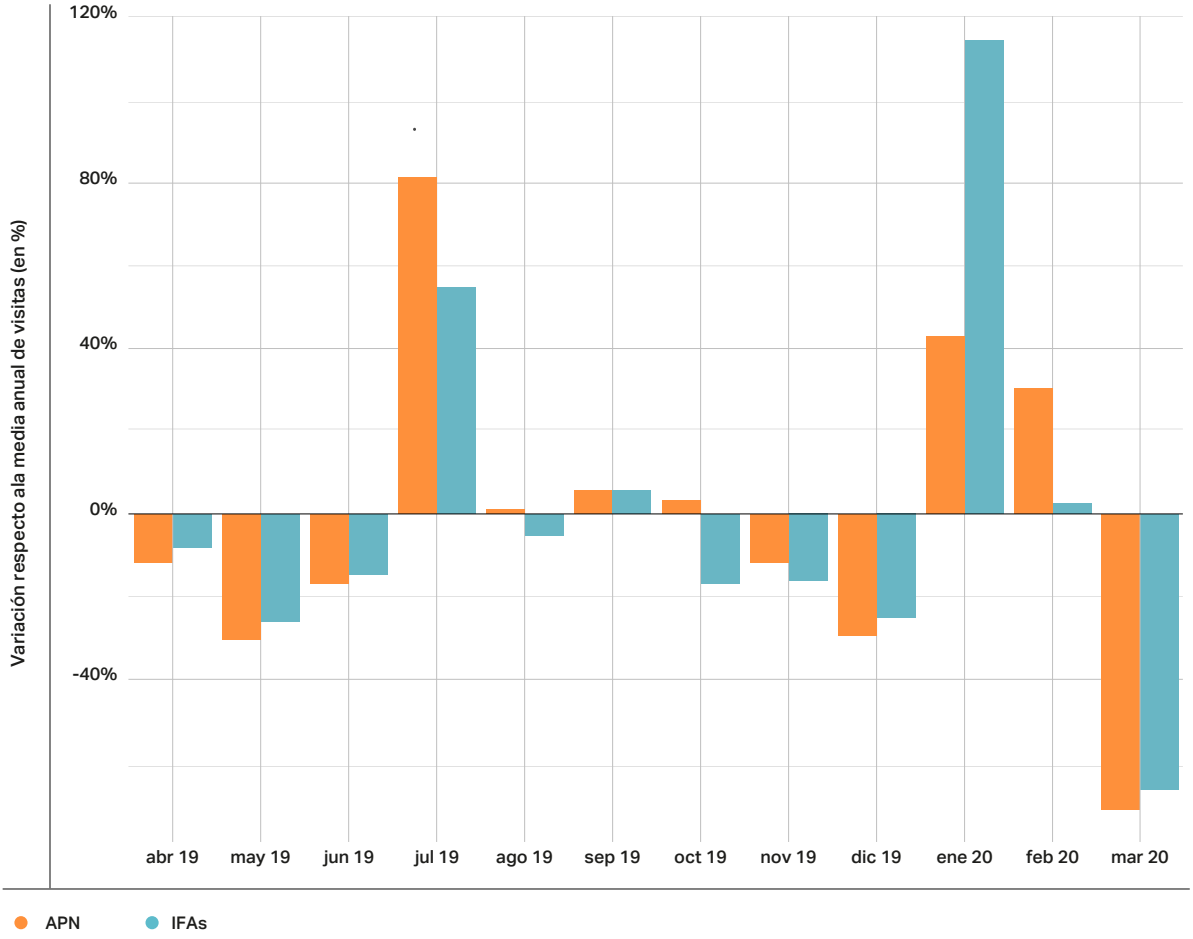
12 Actualmente, en el país existen 50 áreas protegidas nacionales de distintas categorías (Parques Nacionales, Monumentos Naturales y Reservas Nacionales, Reservas Naturales, Reserva Natural Estricta y Reserva Natural Educativa), de las cuales 36 cuentan con métodos sistemáticos de registro de visitas a través de boletos, boleto electrónico, libro de registro, planilla de registro, entre otros. Los informes de seguimiento se encuentran disponibles en el sitio web de Parques Nacionales.

13 A los fines de este documento, un polígono se entiende como una zona o territorio con un límite bien definido.

Casos de uso de la herramienta

Gráfico 8

Visitas a las Cataratas. Evolución de la variación mensual de visitas al Parque Nacional Iguazú (en porcentaje, respecto de la media anual), según registros de ingresos de Administración de Parques Nacionales (APN) y según big data (IFA), durante el período de referencia (abril 2019 a marzo 2020)



Fuente: Fundar, DNMyE, SEDLab - elaboración propia, de acuerdo a datos propios e información del Registro Nacional de Autorizaciones, Recaudaciones e Infracciones (RENARI) de la Administración de Parques Nacionales (APN).

Tomemos como ejemplo el caso del Parque Nacional Iguazú, a la sazón el área protegida con mayor volumen de visitantes registrados del conjunto para el que poseemos información y un sistema de registración confiable (para entrar al parque es necesario abonar un ticket, que es registrado en un sistema informático).

Más allá de las diferencias en las cantidades estimadas por cada fuente, las variaciones mes a mes presentan un comportamiento análogo.

La determinación de la residencia habitual (a partir de su CEL) nos permite ahondar en dos aspectos de los que adolecen los registros administrativos:

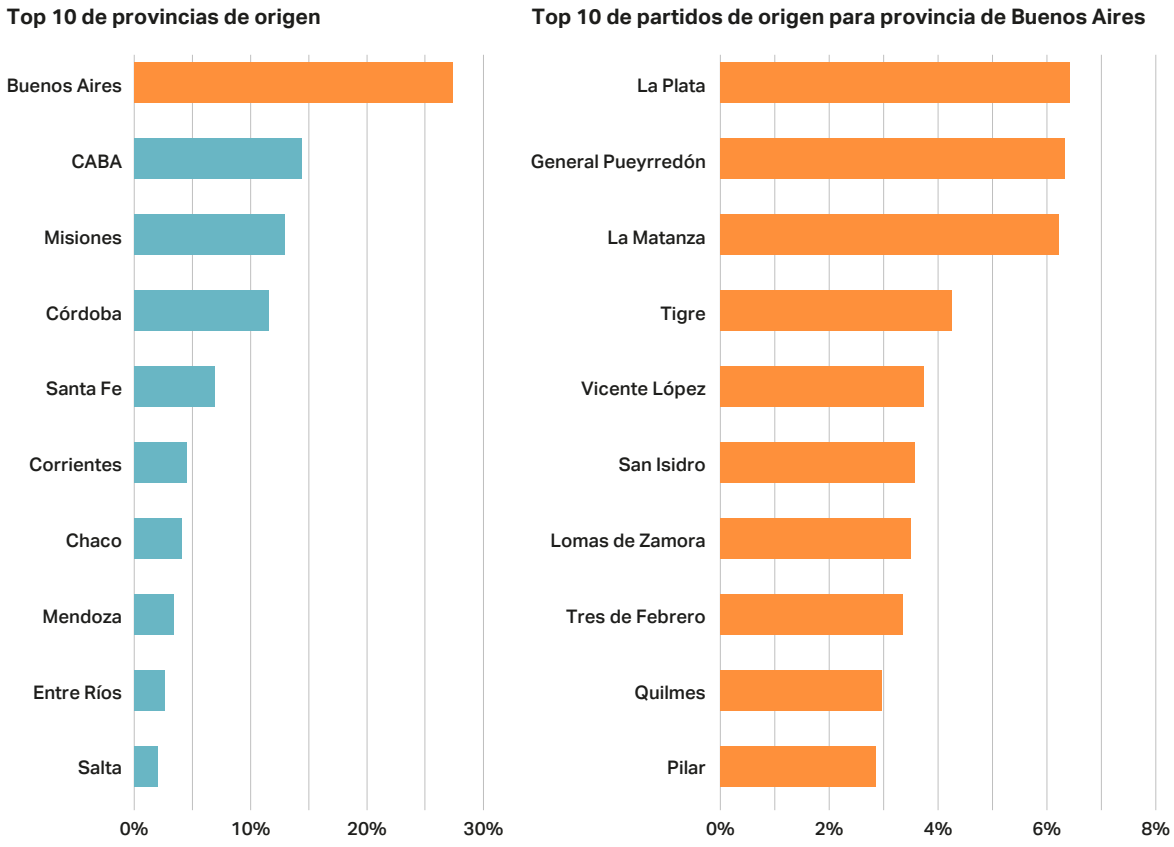
- Origen de los visitantes (ya sea a nivel provincial, por departamento, etc.)
- Nivel socioeconómico de los visitantes.

De esta manera, por ejemplo, podemos ver el *top* de provincias de origen para los visitantes del PN Iguazú, y profundizar el análisis por partido de origen, para el caso de la provincia de Buenos Aires:

Casos de uso de la herramienta

Gráfico 9

Caracterización de los turistas a Cataratas (origen). Ranking de a) las 10 principales provincias y b) los 10 principales municipios de provincia de Buenos Aires de origen de los visitantes (IFA) del Parque Nacional Iguazú

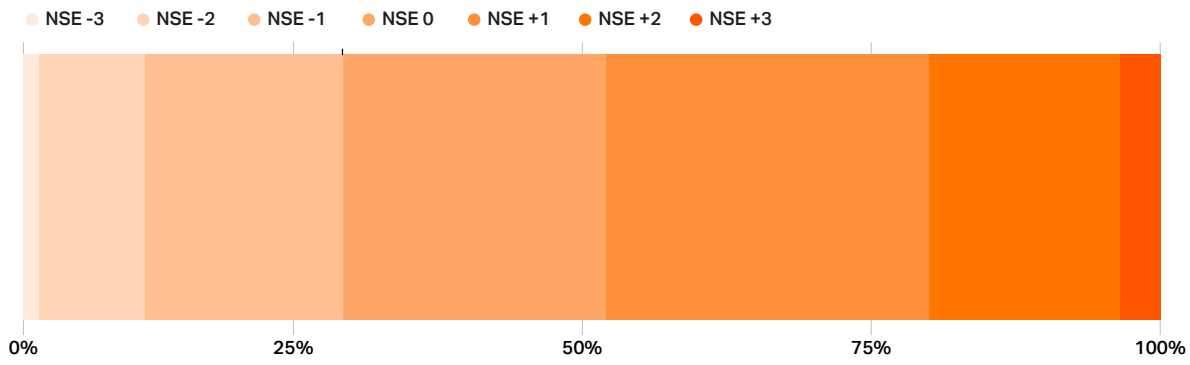


Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

En cuanto al perfil, se ve una mayor proporción de visitantes con un nivel socioeconómico medio-alto (NSE +1, +2, +3) de acuerdo con los registros de IFA.

Gráfico 10

Caracterización de los visitantes a Cataratas (Nivel socioeconómico). Distribución de los visitantes (IFA) al Parque Nacional Iguazú de acuerdo a su nivel socioeconómico (NSE)



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

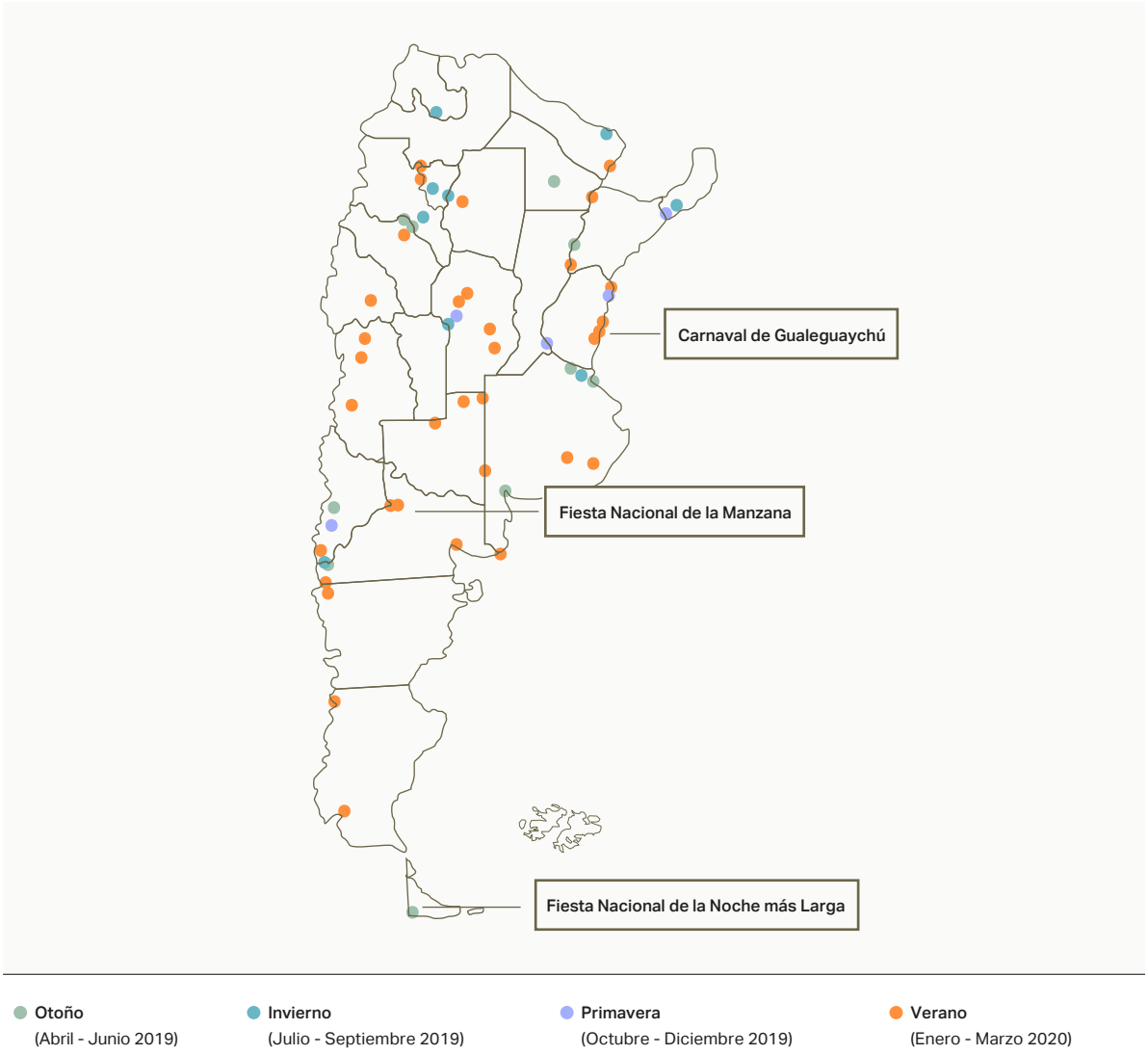
Turismo de fiestas

Desde la Dirección Nacional de Fiestas Nacionales y Eventos (DNFNE) del Ministerio de Turismo y Deportes se lleva un registro de Fiestas en todo el país, con una calendarización aproximada de los eventos (pueden variar año a año) y una caracterización del tipo de actividades de cada uno. Por su propia naturaleza no se dispone de información precisa y estandarizada sobre los visitantes de este tipo de eventos.

Aprovechando la oportunidad de la nueva fuente de datos se realizó un ejercicio piloto para avanzar en este sentido. Con la colaboración de la DNFNE se generó una base de datos de fiestas destacadas en todo el territorio nacional sobre las que se determinó las fechas en que sucedieron (entre abril de 2019 y marzo de 2020) y las coordenadas geográficas de la locación principal de cada evento.

Gráfico 11

Mapa de fiestas nacionales. Distribución geográfica de las fiestas nacionales destacadas, según las coordenadas geográficas de la locación principal de cada evento y las fechas en que sucedieron (entre abril de 2019 y marzo de 2020), de acuerdo al Registro de Fiestas

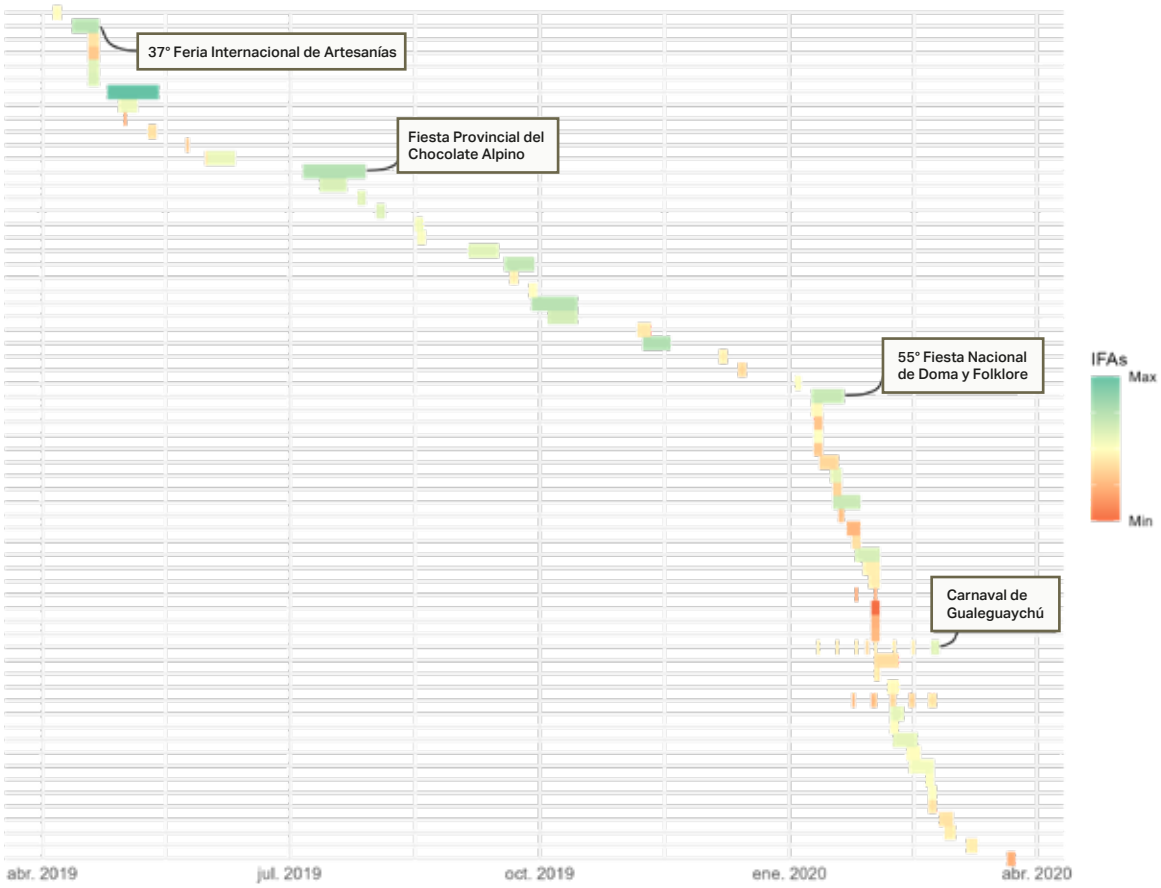


Fuente: Fundar, DNMyE, SEDLab - elaboración propia a partir del registro de la Dirección Nacional de Fiestas Nacionales y Eventos (DNFNE) del Ministerio de Turismo y Deportes.

Casos de uso de la herramienta

Visitas a las fiestas nacionales. Distribución de las fiestas nacionales, según las fechas en que sucedieron (entre abril de 2019 y marzo de 2020) y la cantidad de visitantes (IFA) que recibieron

Gráfico 12



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

A partir de estos datos se generaron consultas a la base de datos GEO y se identificaron los IFA que cumplían con las condiciones para ser considerados visitantes con lo que se obtuvieron señales que residen a más de 20 km del lugar de la fiesta y que en la fecha del evento se hubieran acercado a una distancia menor o igual a 3 km donde este se desarrollaba.

Ejemplo: GUALEGUAYCHÚ

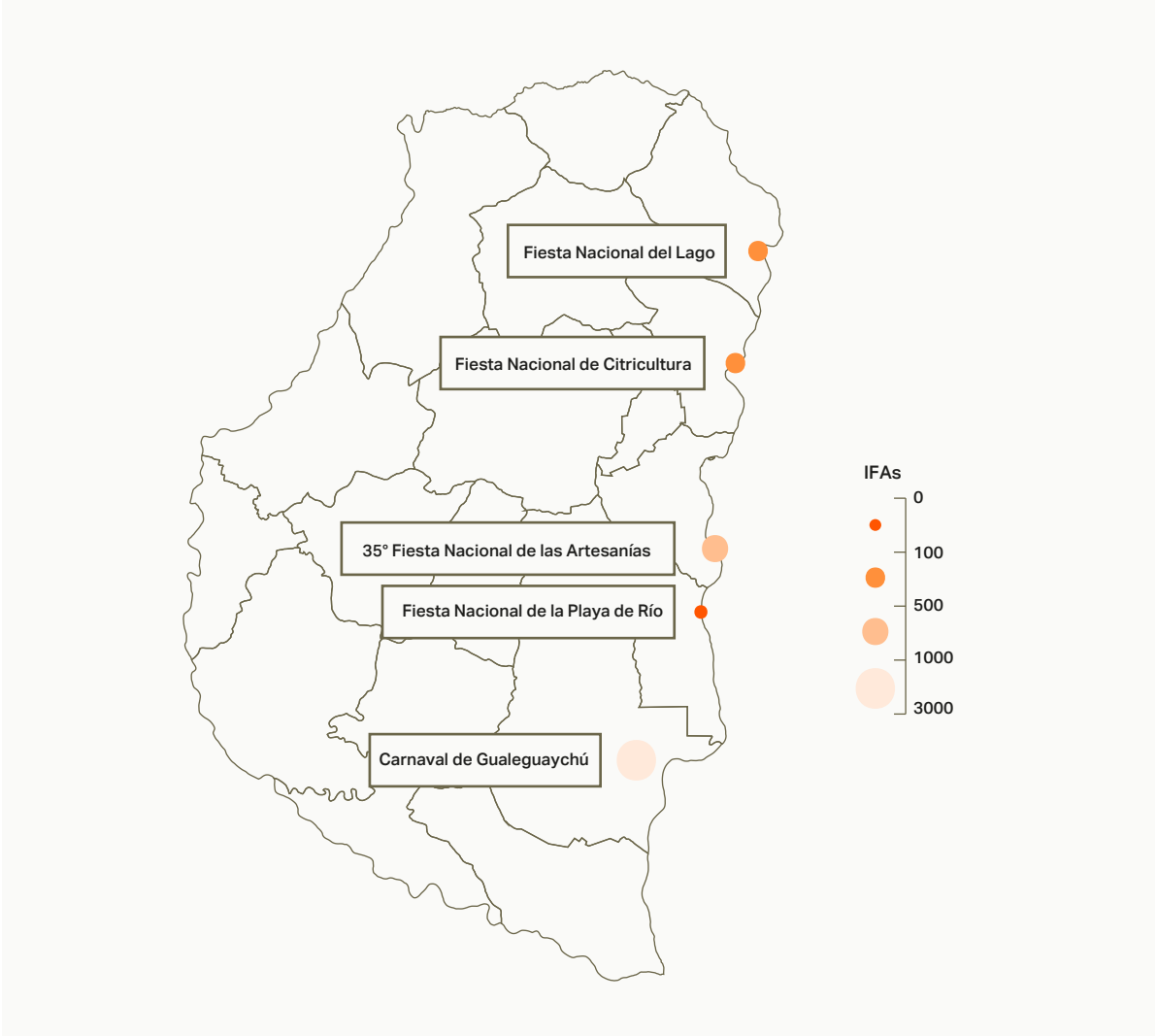
El Carnaval del País, más conocido como el Carnaval de Gualeguaychú, es un evento que se lleva adelante durante los fines de semana de la temporada estival (sábados de enero y febrero) que culmina con el "Feriado de Carnaval" en la ciudad de Gualeguaychú, provincia de Entre Ríos. La identificación de los IFA presentes en determinado momento del tiempo (fechas de carnavales) en un lugar preciso (una distancia menor a los 3 km de Gualeguaychú) nos permite estimar el perfil de los visitantes.

Gráfico 13

Casos de uso de la herramienta

Concurrencia a fiestas nacionales entrerrianas. Mapa de fiestas nacionales de la provincia de Entre Ríos según las coordenadas geográficas de la locación principal de cada evento y cantidades de visitantes (IFA)

Gráfico 13



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

A diferencia de los análisis del turismo interno en general y la comparación con la EVyTH, o la validación de los visitantes de Parques Nacionales contra los datos obtenidos a partir de registros administrativos, no contamos con información análoga para este tipo de eventos. No obstante, podemos valernos de otras fuentes que pueden funcionar como aproximación de lo que estamos analizando.

Tabla 2

Casos de uso de la herramienta

Turismo de fiestas. Ranking de las 10 fiestas nacionales más concurridas de acuerdo a la cantidad de visitas diarias promedio, según IFA

Tabla 2

	Fiestas nacionales	Visitas (cantidad)	Duración (cantidad de días)	Visitas diarias promedio
#1	45° Feria Internacional del Libro	45.205	20	2260
#2	Fiesta Nacional de las Colectividades	5306	10	531
#3	Fiesta Nacional de la Nieve	2564	6	427
#4	37º Feria Internacional de Artesanías	4006	10	401
#5	Fiesta Nacional del Canasto	1152	3	384
#6	49° Fiesta Nacional de la Masa Vienesa	1386	4	346
#7	Fiesta Nacional de los Estudiantes	3160	11	287
#8	Fiesta Nacional del Chocolate	1149	4	287
#9	Carnaval de Gualeguaychú	2779	10	278
#10	Fiesta Nacional Del Chamamé	2364	10	236

Fuente: Fuente: Fundar, DNMyE, SEDLab - elaboración propia a partir del registro de la Dirección Nacional de Fiestas Nacionales y Eventos (DNFNE) del Ministerio de Turismo y Deportes.

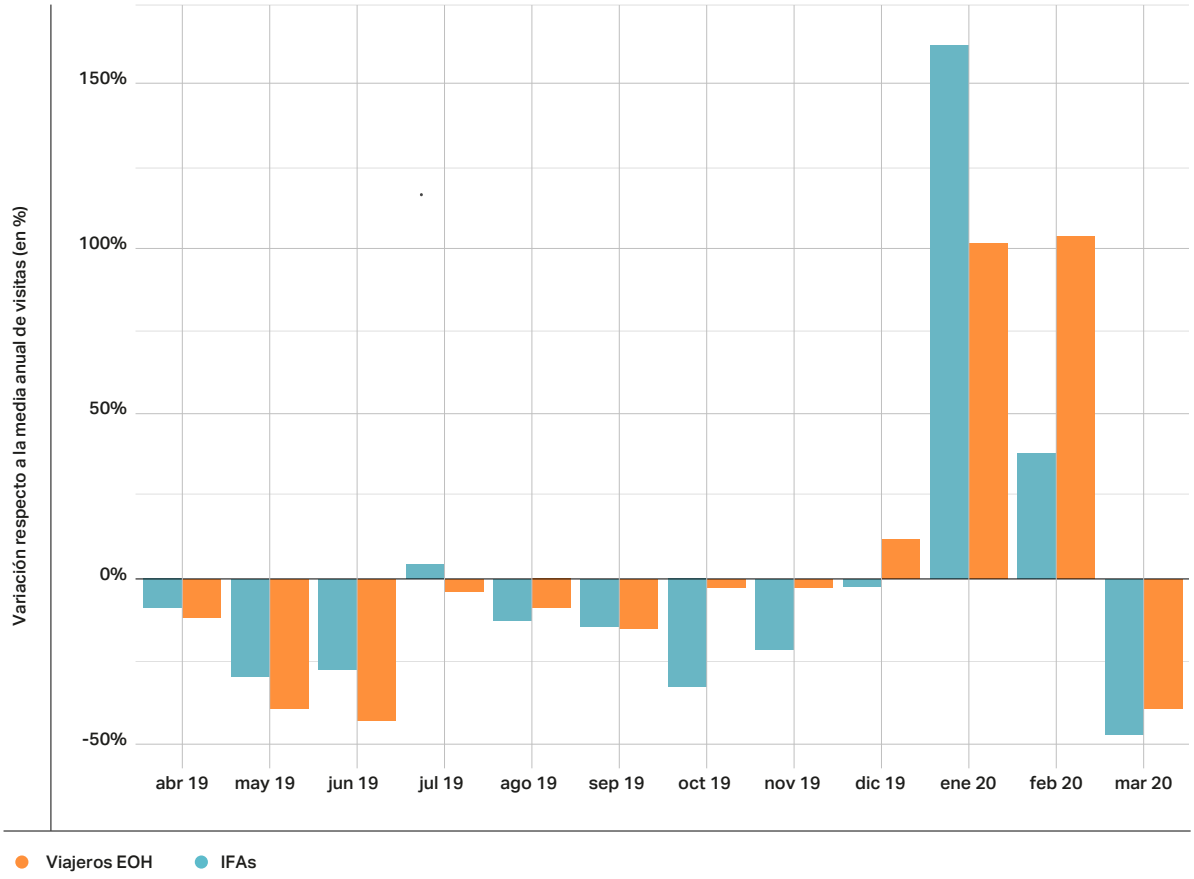
La elección de Gualeguaychú como ejemplo del caso de uso de Turismo en fiestas no es caprichosa: además de ser una de las que concentran mayor cantidad de registros, cuando se ordena al conjunto de fiestas según la variedad de departamentos (domicilio) de procedencia (tal como se describe en la tabla adjunta), la inclusión de la ciudad en la muestra de la Encuesta de Ocupación Hotelera (EOH) nos da la oportunidad de revisar el comportamiento turístico de la nueva fuente de datos *vis a vis* el operativo estadístico.

Gráfico 14

Casos de uso de la herramienta

Visitas anuales a Gualeguaychú. Evolución de la variación mensual de visitas a la ciudad de Gualeguaychú (en porcentaje, respecto de la media anual), según registros de Encuesta de Ocupación Hotelera (EOH) y big data (IFA), durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 14



Fuente: Fundar, DNMyE, SEDLab - elaboración propia , de acuerdo a datos propios e información de la Encuesta de Ocupación Hotelera (EOH), de la Dirección Nacional de Mercados y Estadística (DNMyE) del Ministerio de Turismo y Deportes de la Nación.

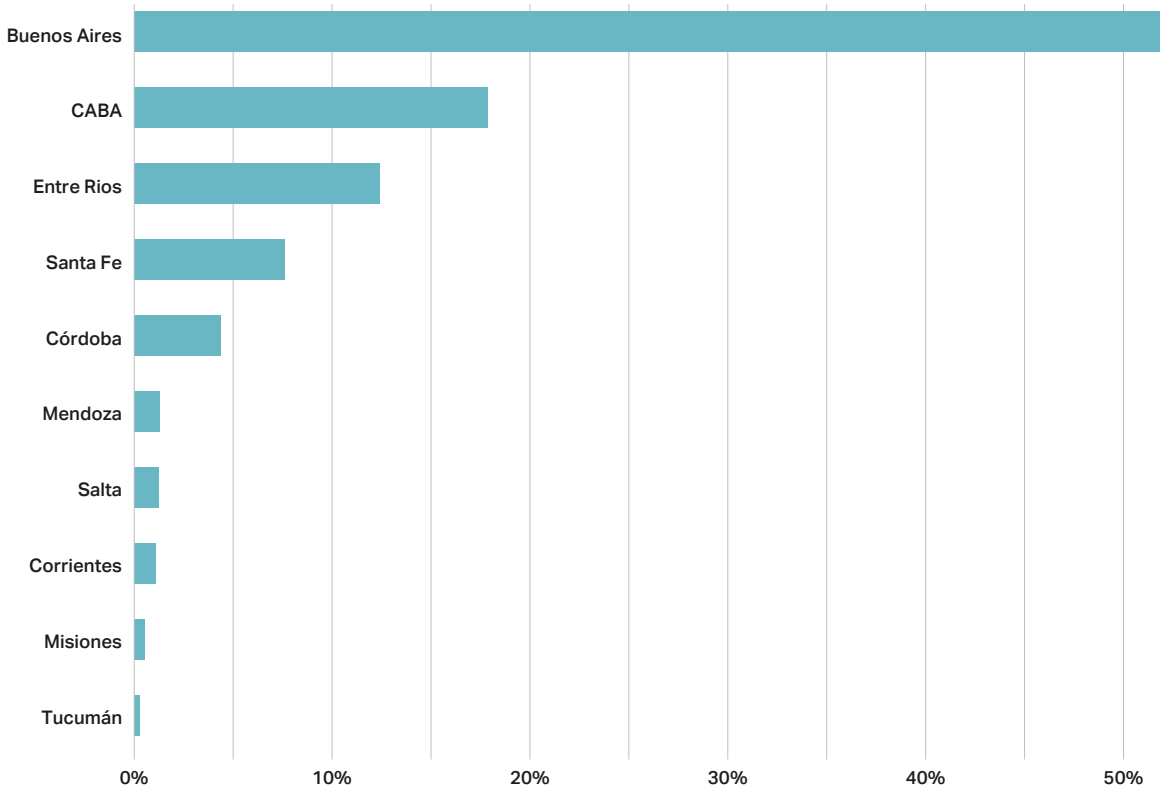
Nuevamente, el ejercicio de identificación de los IFA que, en este caso, participaron de una fiesta, nos permite caracterizar a sus visitantes . Por ejemplo, cuantificando la participación no ya solamente de las provincias de origen - como registra EOH-, sino con mayor desagregación llegando hasta el radio censal y, por ende, podemos observar diferencias socioeconómicas y evaluar perfiles.

Gráfico 15

Casos de uso de la herramienta

Caracterización de los visitantes al Carnaval de Gualeguaychú (provincia de origen). Ranking de las 10 principales provincias de origen de los visitantes del Carnaval de Gualeguaychú (IFA)

Gráfico 15



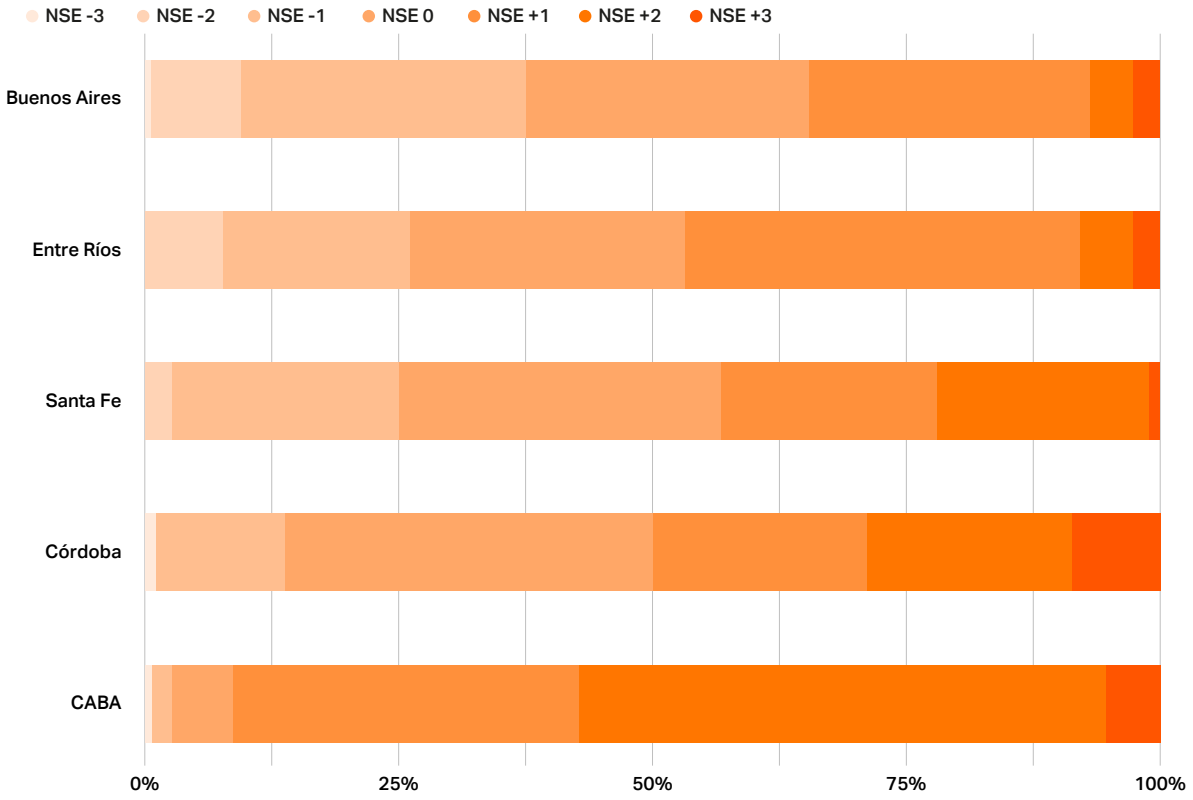
Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Gráfico 16

Casos de uso de la herramienta

Caracterización de los visitantes al Carnaval de Gualeguaychú (Nivel socioeconomico). Distribución de los visitantes al Carnaval de Gualeguaychú (IFA) de las 5 principales provincias de origen, de acuerdo a su nivel socioeconómico (NSE)

Gráfico 16



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

El desarrollo de esta herramienta y metodología de trabajo nos da la posibilidad de repetir este ejercicio para otros casos análogos, por ejemplo el Festival Nacional de Folklore (Córdoba) antes mencionado, y obtener información que no existía previamente a partir de las fuentes de consulta disponibles.

Mercados e intereses: una exploración a través del uso de apps

Aprovechando el hecho de que las geolocalizaciones se realizaron a través de distintas aplicaciones, nos propusimos analizar también su uso con el objetivo principal de caracterizar a los turistas que visitan ciertos tipos de lugares.

La base de datos APPS otorga para cada registro un campo denominado ***Bundleld***, un identificador único de las apps móviles que permite descargar información de las aplicaciones a través de las páginas de Play Store y Apksos. Si bien no pudieron obtenerse los datos de todas, ya que muchas no se encontraban en estos sitios, pudimos recolectar información de 130.000 apps (61% del total) que representan el 90% de los registros. Detectamos que muchas tenían muy pocos registros de geolocalización, por esto nos quedamos con aquellas que tenían más del 1%, es decir, un total de 24.000 apps que representaron el 86,4% de los registros, es decir 1.855.739.415 geolocalizaciones.

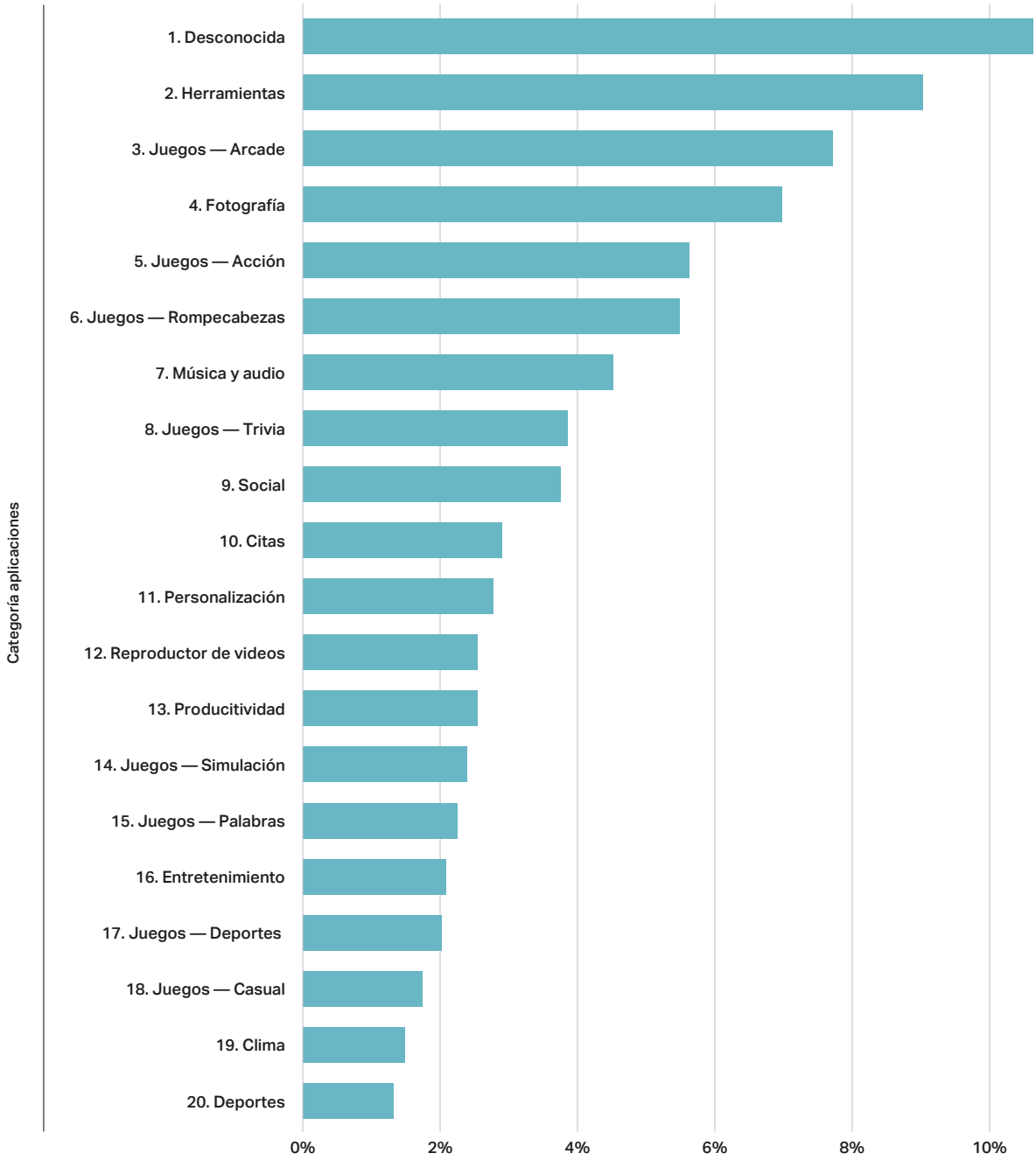
Casos de uso de la herramienta

IFA que viajaron vs. IFA que no

La primera pregunta que nos planteamos sobre estos datos fue si los usuarios que viajaron usaron distintas apps que aquellos que no lo hicieron . En los siguientes gráficos se pueden ver las categorías de apps más usadas por los IFA que viajaron y los que no lo hicieron.

Gráfico 17 a

Uso de aplicaciones de celular de turistas locales. Ranking de las 20 aplicaciones más utilizadas (porcentaje de uso) por turistas (IFA) que viajaron durante el período de referencia (abril 2019 a marzo 2020)

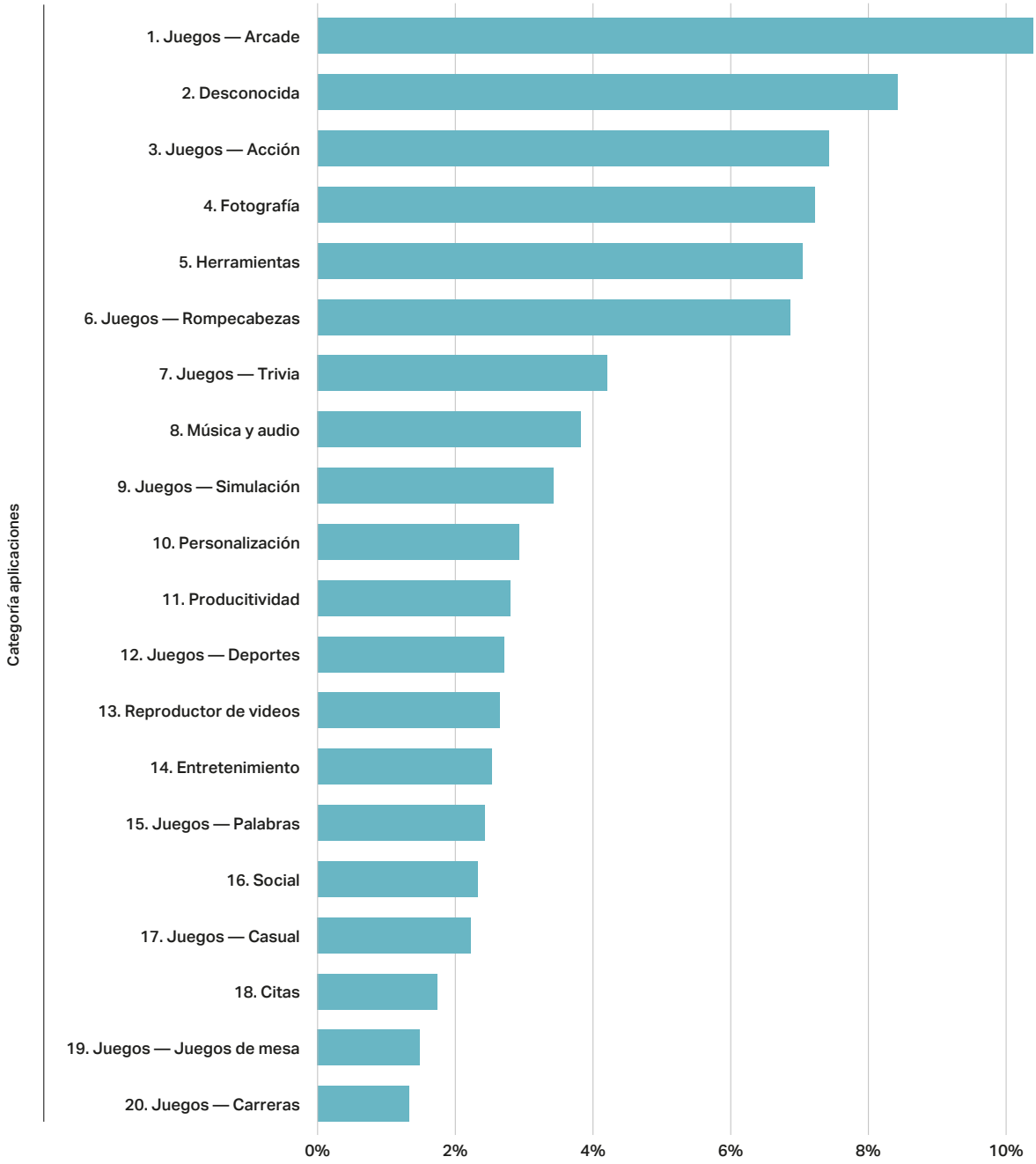


Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

Uso de aplicaciones de celular de no turistas. Ranking de las 20 aplicaciones más utilizadas (porcentaje de uso) por no turistas (IFA) durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 17 b



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

De aquí pudimos sacar algunas observaciones:

- Las apps de tipo juego son las usadas con más frecuencia en ambos casos.
- Los IFA que viajaron usaron más las apps de tipo tools, social y dating.
- También se observa un mayor uso de las apps de tipo clima y sports

Estas observaciones pueden resultar útiles a la hora de definir en qué tipo de apps invertir para acciones de *marketing* en base a si se quiere llegar al público que suele viajar o al que no.

La siguiente pregunta que nos hicimos fue si podíamos estimar el nivel socioeconómico según los tipos de apps, pero esto no arrojó ningún resultado significativo, como tampoco lo tuvimos al intentar ver si había diferencias en las distribuciones de destinos por departamento en base al tipo de app utilizada. Donde sí obtuvimos resultados indicativos fue al analizar casos puntuales de apps y destinos.

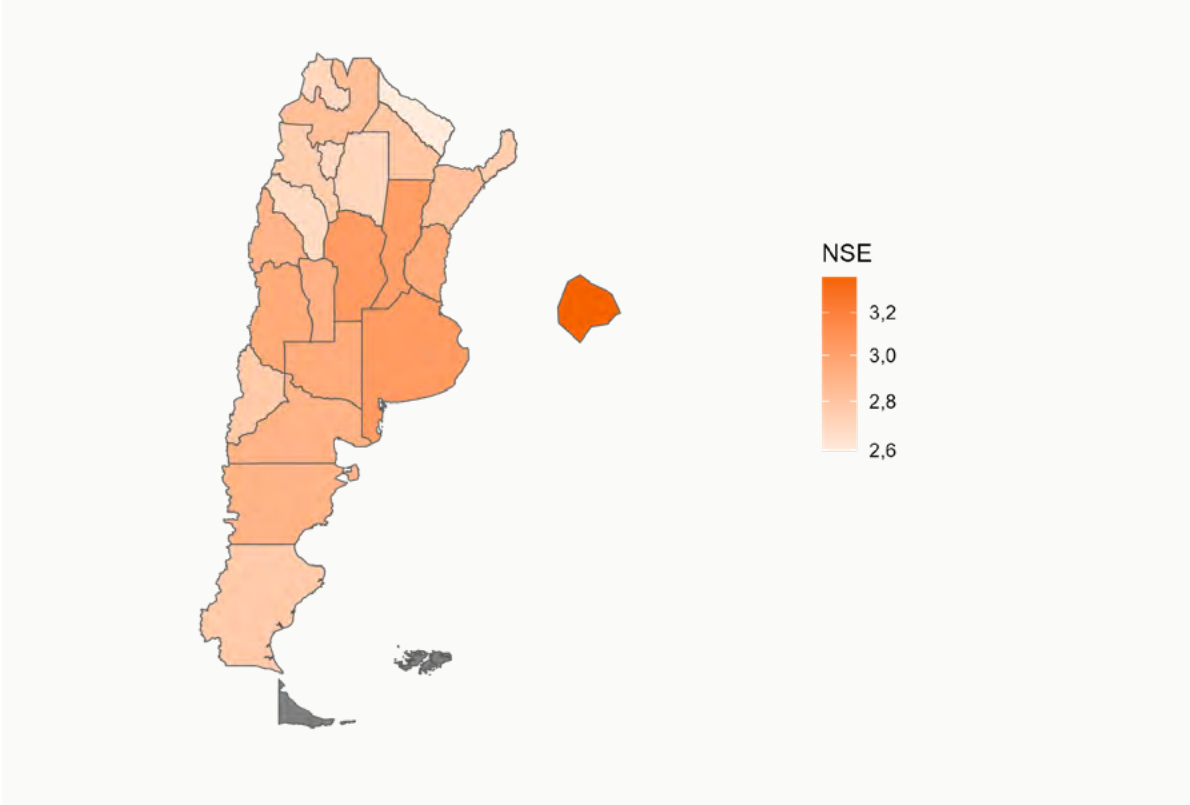
Analizando por app: ejemplo de La Liga Fútbol

Decidimos analizar los casos de los IFA que usaron la app "La Liga Fútbol", que cuenta con 108.000 usuarios.

Al comparar las zonas de residencia cruzadas con los datos del censo 2010 de esta muestra en el siguiente mapa se puede ver el NSE de los IFA de La Liga Fútbol.

Caracterización de los usuarios de La Liga Fútbol. Distribución de los usuarios (IFA) de la aplicación La Liga Fútbol, de acuerdo a su provincia de origen, según su nivel socioeconómico (NSE)

Gráfico 18



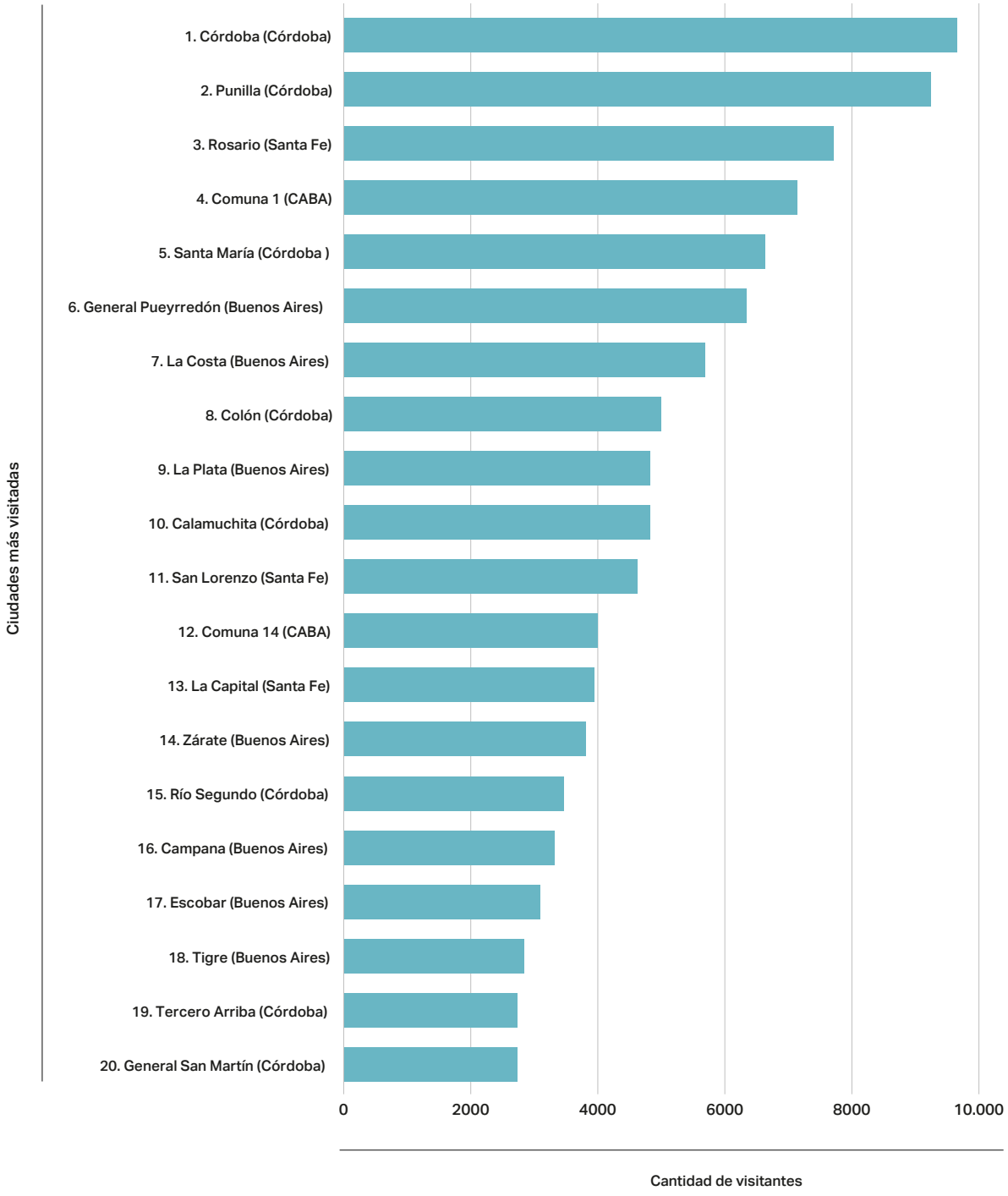
Fuente: F undar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

También quisimos conocer los destinos elegidos por los usuarios de la app; en el siguiente gráfico pueden verse los 20 principales destinos.

Ranking de destinos turísticos más elegidos por los usuarios de La Liga Fútbol. Ranking de las 20 ciudades argentinas más visitadas por los usuarios (IFA) de la aplicación La Liga Fútbol, durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 19



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

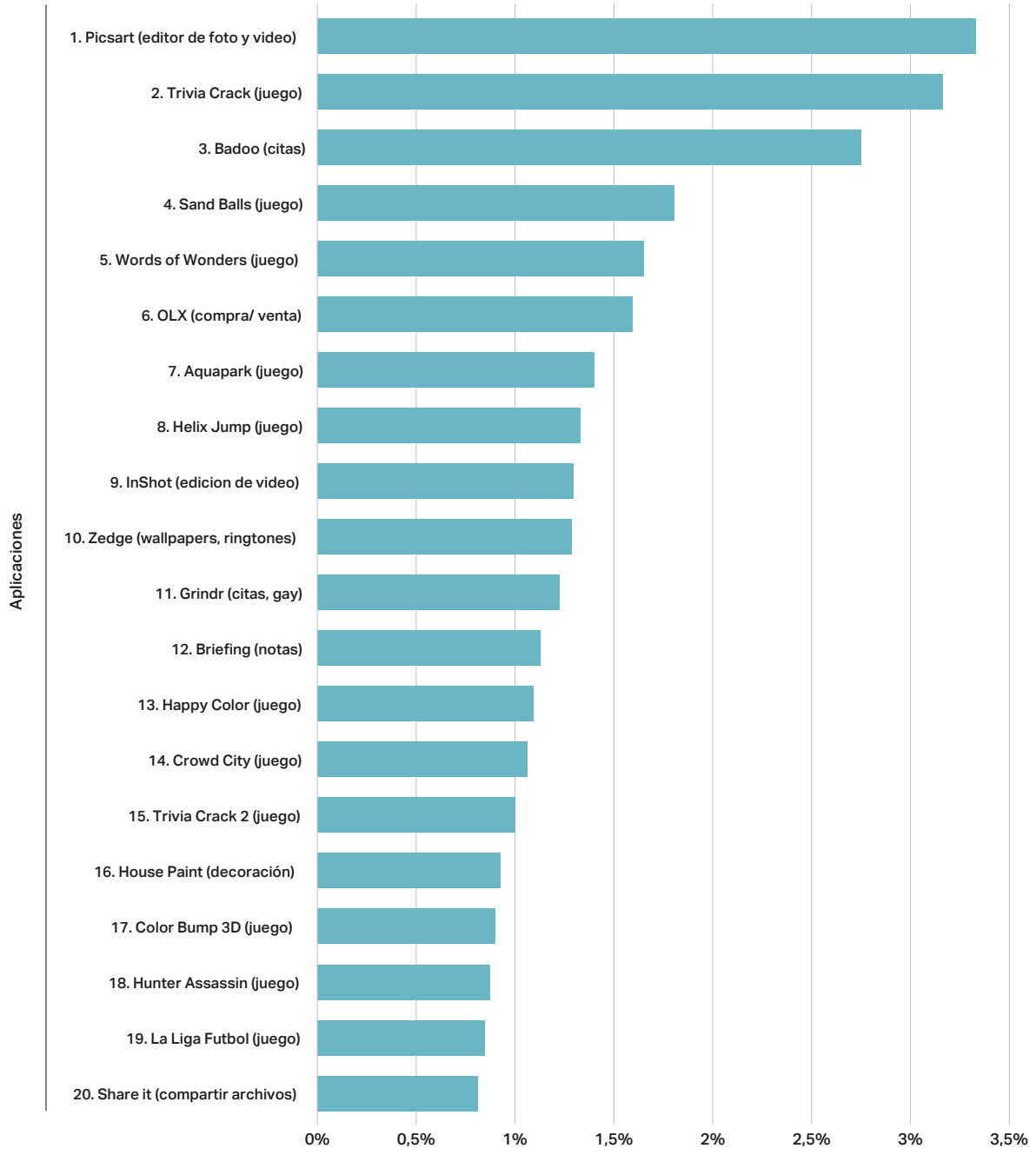
Aquí pudimos ver que los usuarios parecen tener una mayor afinidad por destinos de la provincia de Córdoba. Esta información puede servir a la hora de pautar para esos perfiles específicos de usuarios, para determinados productos turísticos y para pensar estrategias de *marketing* para perfiles y destinos.

Esto nos hizo preguntarnos también si no habría particularidades en otros distritos, por lo que decidimos analizar los casos de Iguazú (todo el año) y Gualeguaychú (durante febrero).

En este caso analizamos las 20 apps más usadas en cada destino normalizando por el total y comparándolo contra el uso de apps de todos los IFA que viajaron (no sólo los que fueron a estos destinos):

Gráfico 20 a

Uso de aplicaciones de celular en Argentina. Ranking de las 20 aplicaciones más utilizadas (porcentaje de uso) por los argentinos (IFA) durante el período de referencia (abril 2019 a marzo 2020)

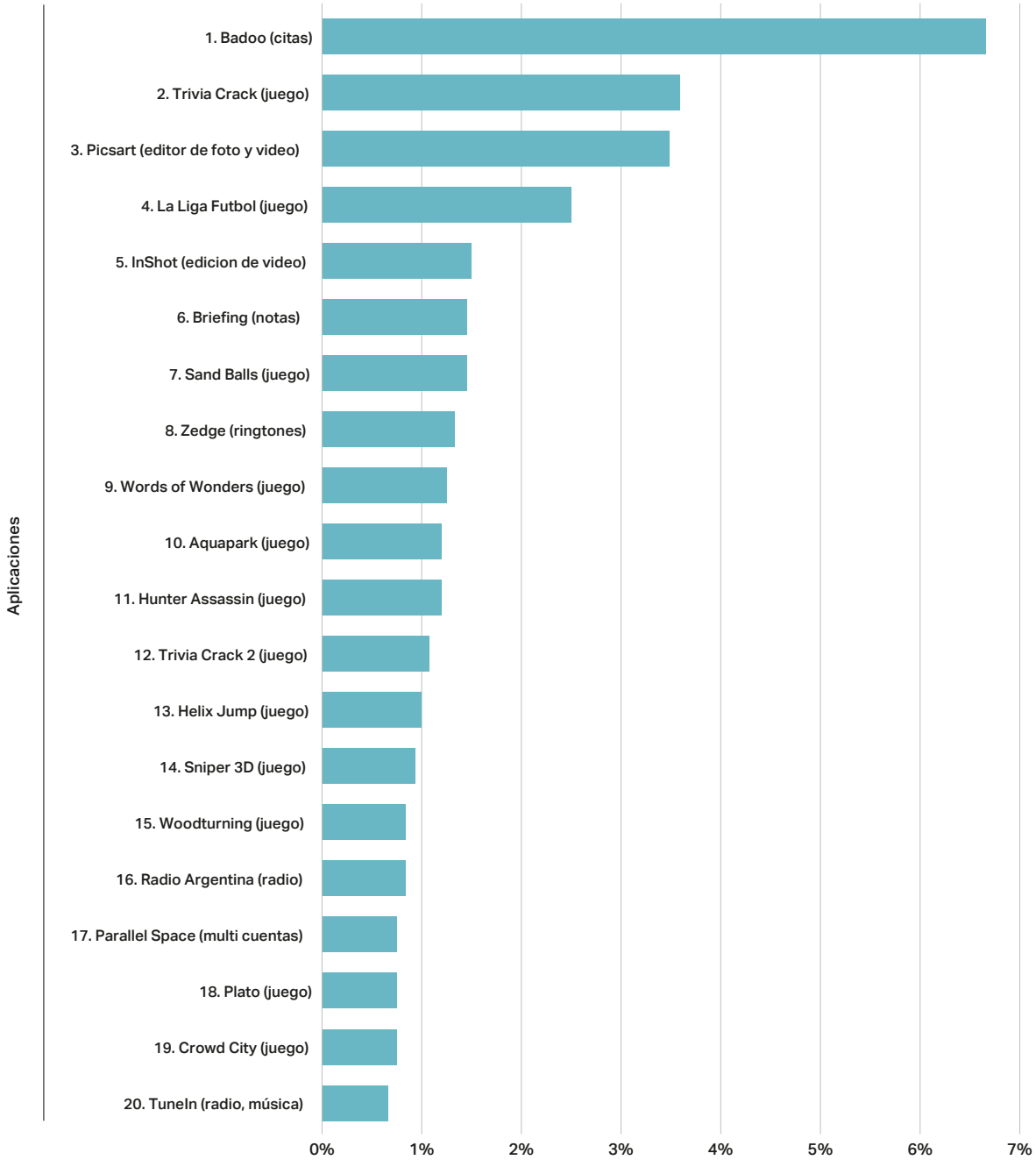


Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Casos de uso de la herramienta

Uso de aplicaciones de celular en Gualeguaychú. Ranking de las 20 aplicaciones más utilizadas (porcentaje de uso) por las personas que se encontraban en la ciudad de Gualeguaychú (IFA) durante el período de referencia (abril 2019 a marzo 2020)

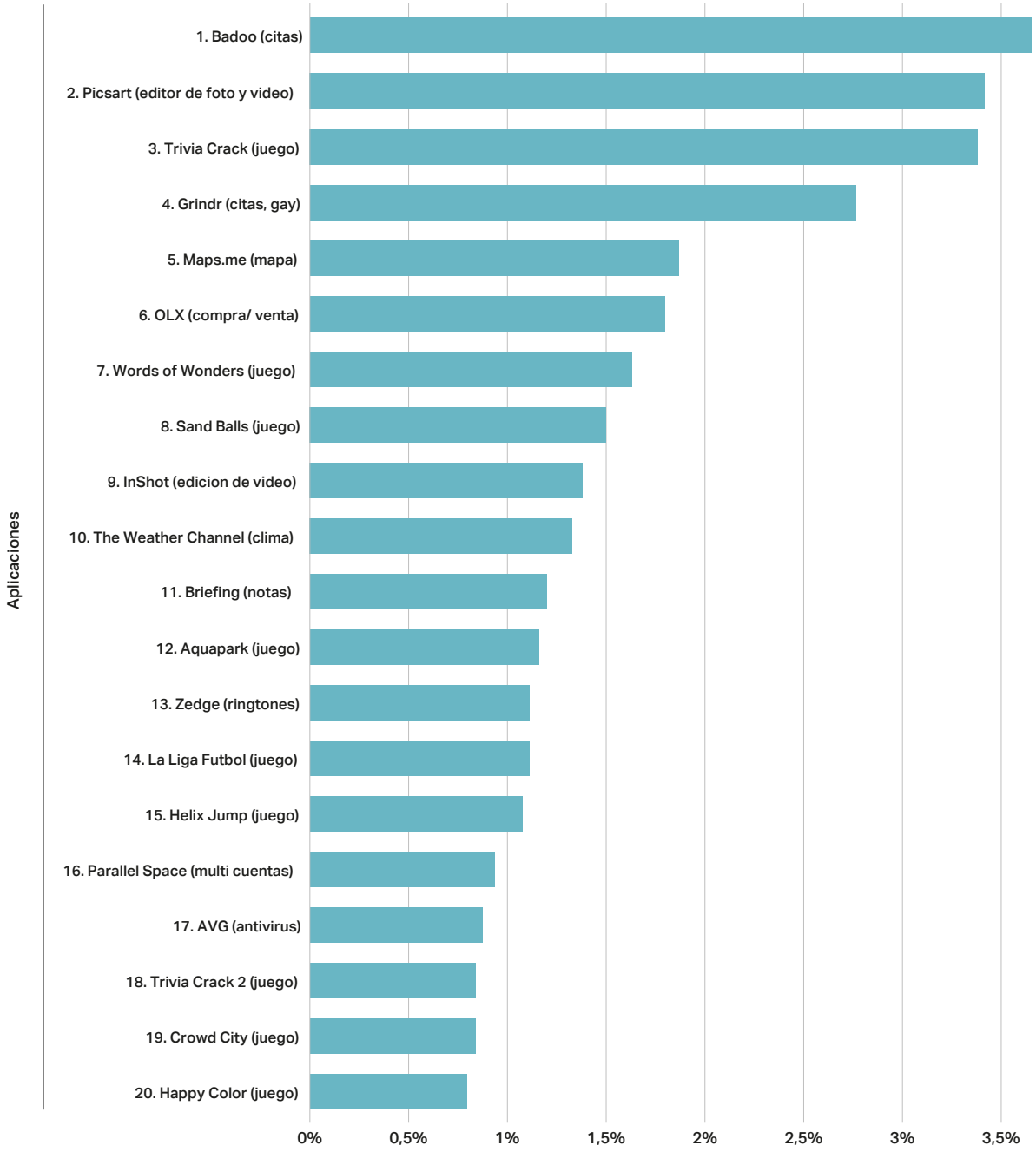
Gráfico 20 b



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Uso de aplicaciones de celular en Iguazú. Ranking de las 20 aplicaciones más utilizadas (porcentaje de uso) por las personas que se encontraban en la ciudad de Iguazú (IFA) durante el período de referencia (abril 2019 a marzo 2020)

Gráfico 20 c



Fuente: Fundar, DNMyE, SEDLab - elaboración propia.

Conclusiones

En estos últimos gráficos se pueden observar varios puntos interesantes. En el caso de Iguazú, Grindr figura entre las 4 apps más utilizadas, mientras que en Gualeguaychú no está ni siquiera entre las 20 más usadas. También vemos que en Iguazú aparecen apps de clima y mapas y en Gualeguaychu no. Por otro lado, en Gualeguaychú la app de fútbol aparece en cuarto lugar y en Iguazú en el puesto 14. Finalmente, Badoo es la app más usada en ambos casos pero en el general es la tercera.

En definitiva, más allá de las limitaciones con las que nos encontramos para utilizar esta fuente de datos, hallamos que pueden proveer información útil para caracterizar o diferenciar a aquellas personas que eligen estos destinos y pensar políticas o campañas para este público en particular.

Conclusiones

Este documento muestra una experiencia práctica de incorporación de fuentes de datos alternativas como complemento a fuentes tradicionales, como encuestas y registros administrativos, haciendo foco en el sector turístico. Las fuentes de datos alternativas, como todas aquellas contenidas en el gran paraguas de "big data", a las cuales se accede mediante instrumentos de procesamiento complejos, son de gran ayuda para conocer a la población a quien el Estado le habla o con quienes se relaciona. Por el nivel de granularidad, aportan conclusiones que muchas veces las fuentes tradicionales no pueden observar.

La fuente alternativa usada en este trabajo es una base de datos georreferenciada que se recopila a partir de dispositivos móviles y brinda una oportunidad para explorar datos que otras fuentes no contemplan. Aporta información a nivel de radios censales, para todo el territorio nacional y con granularidad temporal diaria. Para el sector turístico este tipo de ejercicio brinda una oportunidad para explorar información de origen de los viajes al nivel de radios censales y referencia desagregada sobre los destinos de esos viajes. A su vez, en algunos casos permite analizar ciertos recorridos realizados para desplazarse hacia los destinos, lo que complementa la visión puramente cuantitativa.

Los datos de las fuentes de datos alternativas no provienen de un diseño experimental y su tamaño y organización son dispersos: la cantidad de datos no es pareja ni espacial ni temporalmente. Por lo tanto, su calidad estadística es desconocida y los datos requieren depuración. En ese contexto, , el objetivo es buscar datos que de alguna forma brinden información confiable o de calidad. En este análisis, buscar datos de calidad implicó contar a aquellos que tienen definida una "common evening location" (CEL), que es la ubicación típica durante la noche de ese dispositivo. Se asume como criterio que esa ubicación típica es la residencia de la persona. De esta forma se pueden identificar los datos a utilizar en el análisis, eliminando cuestiones irrelevantes o que introducen ruido, y generar una base de datos novedosa por el nivel de cobertura y desagregación territorial (información al nivel de radios censales de todos los departamentos de las 24 jurisdicciones).

Llevar a cabo este tipo de análisis sobre una base de datos con información de personas permite distinguir grupos y perfiles distintos basados en las variables con las que se cuenta. Esta herramienta fue y es muy utilizada en *marketing* o comercialización para la segmentación de clientes. En este trabajo se propone difundir su uso para analizar y conocer en mayor profundidad a la población objetivo de una política pública o para otros casos como el de la evaluación de impacto de potenciales políticas mediante modelos de simulación. En el sector turístico también se pueden comparar estos datos con la información de turistas residentes en pasos migratorios terrestres, puertos y aeropuertos, analizar eventos puntuales (deportivos, musicales, etc.) o actividades específicas (turismo de nieve), evaluar la movilidad inter e intra destinos o medir las distancias recorridas.

La falta de datos impacta en el diseño de políticas públicas, la toma de decisiones de gestión y la evaluación de impacto de las políticas. El uso de fuentes alternativas de información como complemento a bases de datos tradicionales ofrece un camino, no siempre perfecto, de remediación. Estas fuentes pueden tener múltiples orígenes y, como vimos en el trabajo, combinarse en forma exitosa y a la vez generar nuevos desafíos.

Metodología

En este documento se presentó un ejercicio de uso de fuentes de datos alternativos junto con información ya conocida usada como ancla. Así, los casos de uso se valieron de recursos como la Encuesta de Ocupación Hotelera (EOH) al analizar el turismo de fiestas en Gualeguaychú, o los registros de visitas generados por la Administración de Parques Nacionales (APN) cuando analizamos los registros en el área del Parque Nacional Iguazú. También buscamos validar la calidad de la información en relación con la Encuesta de Viajes y Turismo de los Hogares (EVyTH), el operativo estadístico de base para analizar el turismo interno en la Argentina, en el que observamos un alto grado de correlación con las variaciones a lo largo del tiempo o las estimaciones de rankings de destinos.

Si bien todos estos ejercicios fueron de utilidad para poner en contexto la nueva información que procesamos, también nos marcaron limitaciones. Por ejemplo, la EVyTH cuenta en su diseño con un instrumento central para generar indicadores clave de seguimiento del turismo, que es el "Módulo de Comportamiento Turístico". A través de él se puede estimar (para la muestra del operativo) la proporción de la población que realizó viajes turísticos en un año dado. A pesar de que las ventanas de información difieren, nos propusimos calcular ese mismo indicador para la muestra de IFA con residencia estimada (CEL) dentro de los GAU del Indec. El resultado no se aproximó como esperábamos: calculamos casi 20 puntos porcentuales de diferencia entre EVyTH e IFA, 38% vs. 18%, respectivamente.

Pueden hacerse algunas consideraciones sobre aspectos centrales del uso y procesamiento de big data:

- **Calidad**

Partiendo de una cantidad muy grande de datos con estructura y calidad posiblemente muy disímiles y poco conocidas, es posible plantearse preguntas bien definidas para luego quedarse con porciones de los datos que cumplan requisitos mínimos para las preguntas formuladas. Por ejemplo, considerar IFA que tengan su lugar de residencia bien definido y muestren una actividad de movilidad mínima durante toda la ventana de tiempo a ser estudiada.

- **Sesgos**

Ya tenemos datos que creemos sirven para responder nuestras preguntas, pero ¿ocultan sesgos no evidentes? Podemos controlar qué tanto se asemejan nuestros datos a "características globales" de la población y sacar conclusiones. Por ejemplo, es posible relacionar los radios censales de residencia de los IFA y controlar si la proporción de población representada (para diferentes NSE) es comparable con la misma proporción pero en la población general.

- **Filtrado**

Con un subconjunto de datos que ya controlamos desde los puntos de vista de (algunos) aspectos de calidad y sesgo, ya podemos aplicarle "criterios específicos" propios de nuestra disciplina, en este caso, el turismo. Por ejemplo, queremos considerar como "viaje" solo aquellos registros georreferenciados que se ubican a más de 20 km de distancia de su hogar (o 40 km si su residencia está dentro de AMBA). Al resto de los datos los dejamos a mano, pero no los consideramos para el estudio. Este filtrado solo es razonable luego de habernos quedado previamente con datos confiables.

- **Uso de apps**

Utilizando el BundleId es posible obtener datos de las apps de forma gratuita a través del scrapeo de los sitios Google Play y Apksos. Reemplazando en la url el valor de "id" en el primer caso o el último texto después de "/" en el segundo por el BundleId deseado, accederemos al sitio donde está publicada la

Metodología

información de la app (si es que la app existe en dicho store). Esto nos retornará una página web con los datos públicos de la app. Mediante librerías como **BeautifulSoup** podremos seleccionar que parte del HTML deseamos conservar. Finalmente podremos saber qué aplicación se usó en cada registro de geo-localización cruzando el registro con la información obtenidas de las apps a través del BundleId.

Anexo

LO QUE SABEMOS GRACIAS AL **BIG DATA**

VISITANTES DEL PARQUE NACIONAL IGUAZÚ

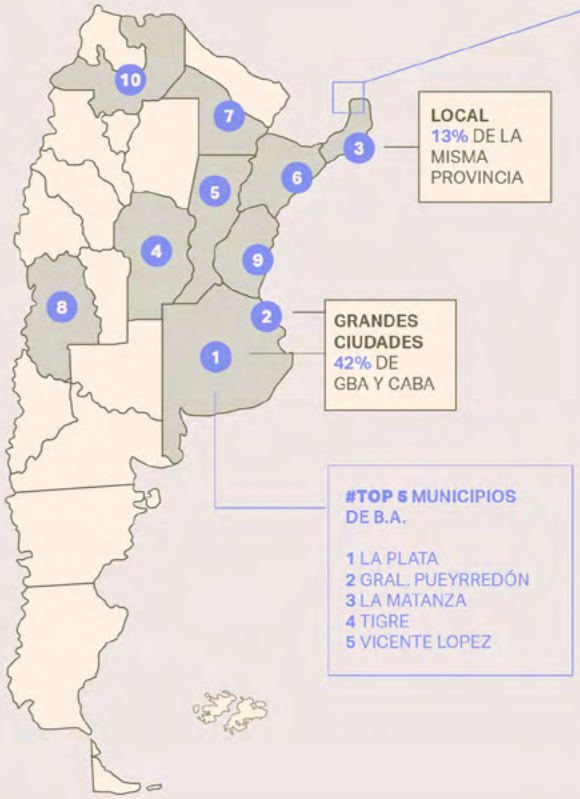
MISIONES, ARGENTINA

Turismo de naturaleza

DESDE DÓNDE

1

#TOP 10 TURISTAS EN CATARATAS POR PROVINCIA



LOCAL
13% DE LA MISMA PROVINCIA

GRANDES CIUDADES
42% DE GBA Y CABA

#TOP 5 MUNICIPIOS DE B.A.

1 LA PLATA
2 GRAL. PUEYRREDÓN
3 LA MATANZA
4 TIGRE
5 VICENTE LOPEZ

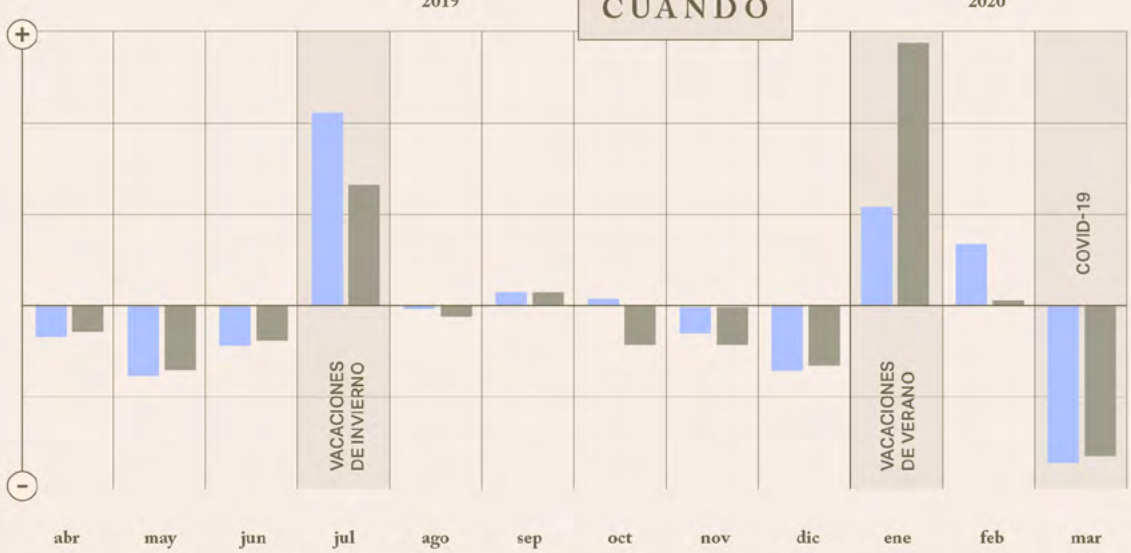
QUIÉNES

2

NIVEL SOCIOECONÓMICO DE LOS VISITANTES A CATARATAS

CUÁNDO

3



VISITAS A CATARATAS

VARIACIÓN MENSUAL RESPECTO DE LA MEDIA ANUAL

Referencias

SEGÚN DATOS DEL MINISTERIO DE TURISMO

SEGÚN BIG DATA

LO QUE SABEMOS GRACIAS AL **BIG DATA**

VISITANTES DEL CARNAVAL DE GUALEGUAYCHÚ

ENTRE RÍOS, ARGENTINA

Turismo de fiestas

1 DESDE DÓNDE

#TOP 5 TURISTAS EN GUALEGUAYCHÚ POR PROVINCIA

LOCAL
15% DE LA MISMA PROVINCIA

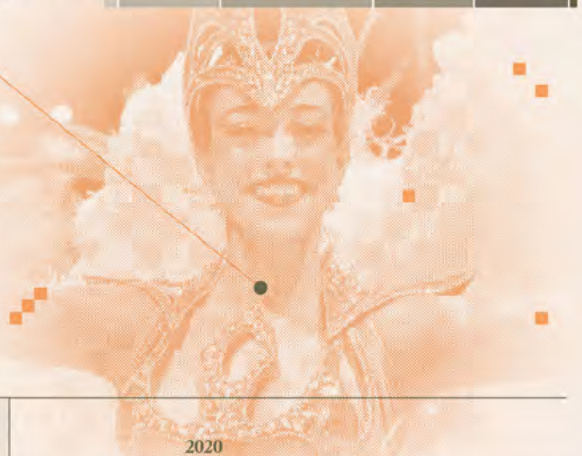
GRANDES CIUDADES
70% DE GBA Y CABA

2 QUIÉNES

NIVEL SOCIOECONÓMICO DE LOS VISITANTES A GUALEGUAYCHÚ SEGÚN PROCEDENCIA

NSE-2 NSE -3 | NSE-1 | NSE 0 | NSE+1 | NSE+2 | NSE+3

1. *Buenos Aires*
2. *CABA*
3. *Entre Ríos*
4. *Santa Fe*
5. *Córdoba*



3 CUÁNDO

VISITAS A GUALE-GUAYCHÚ

VARIACIÓN MENSUAL RESPECTO DE LA MEDIA ANUAL

Referencias

SEGÚN DATOS DEL MINISTERIO DE TURISMO

SEGÚN BIG DATA

Bibliografía

• Brust, A. V., Olego, T., & Rosati, G. (2018). Construcción de un Mapa de Vulnerabilidad Sanitaria en Argentina a partir de datos públicos. Disponible en https://www.fundacionbyb.org/

• D'Alessandro, Mercedes (2022). Ingreso Familiar de emergencia: una política pública a contrarreloj. Buenos Aires: Fundar. Disponible en https://www.fund.ar

• Fundar (2021). La anonimización: un instrumento clave para una gestión de datos eficiente. Buenos Aires. Fundar. Disponible en https://www.fund.ar

• Ministerio de Turismo y Deportes de la Nación (2022). Encuesta de Viajes y Turismo de los Hogares (EVyTH): Documento Metodológico. Subsecretaría de Desarrollo Estratégico. Dirección Nacional de Mercados y Estadística.

• Ministerio de Turismo y Deportes de la Nación (2022). Fuentes de datos: Estimación Nacional del Turismo Internacional y Encuesta de Turismo Internacional (ETI). Subsecretaría de Desarrollo Estratégico. Dirección Nacional de Mercados y Estadística.

• Ministerio de Turismo y Deportes de la Nación (2022).Fuentes de datos: Encuesta de Ocupación Hotelera (EOH). Subsecretaría de Desarrollo Estratégico. Dirección Nacional de Mercados y Estadística.

• Ministerio de Turismo y Deportes de la Nación (2022). Fuentes de datos: Encuesta de Viajes y Turismo de los Hogares (EVyTH). Subsecretaría de Desarrollo Estratégico. Dirección Nacional de Mercados y Estadística.

• Encuesta de Viajes y Turismo de los Hogares (EVyTH).

• Yankelevich, Daniel (2021). Anónimos pero no tanto: cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad. Buenos Aires: Fundar. Disponible en https://www.fund.ar/

Acerca del equipo autoral

Daniel Yankelevich
Director del Área de Datos de Fundar.

Informático, licenciado en la ESLAI en Argentina y recibió su PhD en la Universidad de Pisa. Realizó un postdoctorado en Carolina del Norte, EEUU. Ha realizado tareas docentes y de investigación en varias universidades en Argentina y como profesor invitado en otras universidades de la región.

Mariana Kunst
Coordinadora del Área de Datos de Fundar

Licenciada en Economía y maestranda en Métodos Cuantitativos para la Gestión y Análisis de Datos por la Universidad de Buenos Aires. Se desempeñó como coordinadora del Sistema de Información Cultural de la Argentina (SInCA). Actualmente es docente en la Universidad de Buenos Aires.

Juan Manuel Ortiz de Zárate
Científico de Datos de Fundar

Licenciado y doctorando en Ciencias de la Computación por la Universidad de Buenos Aires. Su tesis de doctorado se enfoca en el estudio de las redes sociales a través del procesamiento del lenguaje natural (NLP) y análisis de grafos. Fue ingeniero de software durante 10 años en empresas de distintos sectores: telecomunicaciones, periodismo, brokers y consultoría política.

Instituto de Cs. de la Computación UBA-CONICET.
Laboratorio de Simulación de Eventos Discretos

Rodrigo Castro

Ing. en Electrónica y Doctor en Ingeniería por la Universidad Nacional de Rosario. Se especializó en simulación de sistemas complejos socionaturales en el Politécnico Federal de Suiza (ETH Zürich). Es director del Laboratorio de Simulación de Eventos Discretos e investigador del Instituto de Cs. de la Computación (ICC UBA-CONICET) y profesor en el Departamento de Computación de la Facultad de Cs. Exactas y Naturales, UBA.

Rafael Grimson

Lic. en Cs. Matemáticas por la Universidad de Buenos Aires, Mg. en Informática por la École Polytechnique de París y Dr. en Ciencias por la Universidad de Hasselt (Bélgica). Se especializó en análisis de datos geoespaciales para problemas ambientales. Es investigador en el Instituto de Investigaciones e Ingeniería Ambiental (UNSAM-CONICET).

Mariano Zapatero

Analista Universitario en Computación por el Departamento de Computación (FCEyN-UBA) y posee una vasta experiencia liderando proyectos de plataformas informáticas para simulación, análisis y visualización de datos, tanto para el sector industrial como el de ciencia y tecnología.

Ministerio de Turismo y Deportes de la Nación

Juan Pablo Ruiz Nicolini

Director Nacional de Mercados y Estadística - Ministerio de Turismo y Deportes de la Nación, magíster en Ciencia Política y Licenciado en Ciencia Política y Gobierno, Universidad Torcuato Di Tella, donde también es docente de Ciencia de Datos en programas de Maestría en Economía Urbana y Maestría en Ciencia Política.

Elián Soutullo

Analista de datos Dirección Nacional de Mercados y Estadística - Ministerio de Turismo y Deportes de la Nación Licenciado en Turismo, Universidad Nacional de La Plata. Se especializa en ciencia de datos aplicada al análisis y desarrollo del turismo.

Juan Gabriel Juara

Coordinador del Área de Datos - Dirección Nacional de Mercados y Estadísitca - Ministerio de Turismo y Deportes de la Nación Licenciado en Sociología, Universidad de Buenos Aires. Cientista de Datos.

Dirección ejecutiva: Martín Reydó

Revisión Institucional: Juliana Arellano

Corrección: Gonzalo Fernández Rozas

Diseño: Jimena Zeitune

Modo de citar

Yankelevich, D.; Soutullo, E.; Juara, J. G.; Ortiz de Zárate, J. M.; Ruiz Nicolini, J. P.; Kunst, M.; Zapatero; M.; Grimson, R. y Castro, R. (2023). De Ushuaia a La Quiaca: byte por byte. ¿Cómo potenciar el sector turístico con big data?. Buenos Aires: Fundar. Disponible en https://www.fund.ar.
DOI: 10.5281/zenodo.8136552

Sobre Fundar

Fundar es un centro de estudios y diseño de políticas públicas que promueve una agenda de desarrollo sustentable e inclusivo para la Argentina. Para enriquecer el debate público es necesario tener un debate interno: por ello lo promovemos en el proceso de elaboración de cualquiera de nuestros documentos. Confiamos en que cada trabajo que publicamos expresa algo de lo que deseamos proyectar y construir para nuestro país. Fundar no es un logo: es una firma.

Trabajamos en tres misiones estratégicas para alcanzar el desarrollo inclusivo y sustentable de la Argentina:

Generar riqueza. La Argentina tiene el potencial de crecer y de elegir cómo hacerlo. Sin crecimiento, no hay horizonte de desarrollo, ni protección social sustentable, ni transformación del Estado. Por eso, nuestra misión es hacer aportes que definan cuál es la mejor manera de crecer para que la Argentina del siglo XXI pueda responder a esos desafíos.

Promover el bienestar. El Estado de Bienestar argentino ha sido un modelo de protección e inclusión social. Nuestra misión es preservar y actualizar ese legado, a través del diseño de políticas públicas inclusivas que sean sustentables. Proteger e incluir a futuro es la mejor manera de reivindicar el espíritu de movilidad social que define a nuestra sociedad.

Transformar el Estado. La mejora de las capacidades estatales es imprescindible para las transformaciones que la Argentina necesita en el camino al desarrollo. Nuestra misión es afrontar la tarea en algunos aspectos fundamentales: el gobierno de datos, el diseño de una nueva gobernanza estatal y la articulación de un derecho administrativo para el siglo XXI.

En Fundar creemos que el lenguaje es un territorio de disputa política y cultural. Por ello, sugerimos que se tengan en cuenta algunos recursos para evitar sesgos excluyentes en el discurso. No imponemos ningún uso en particular ni establecemos ninguna actitud normativa. Entendemos que el lenguaje inclusivo es una forma de ampliar el repertorio lingüístico, es decir una herramienta para que cada persona encuentre la forma más adecuada de expresar sus ideas.

Sobre DNMYE

La Dirección Nacional de Mercados y Estadística de la Subsecretaría de Desarrollo Estratégico del Ministerio de Turismo y Deportes de la Nación forma parte del Sistema Estadístico Nacional (SEN) y se encarga de elaborar, recopilar, interpretar y/o divulgar estadísticas oficiales referidas al turismo.

Sistema de Información Turística de la Argentina (SINTA)
https://www.yvera.tur.ar/sinta/

Sobre SEDLab

El Laboratorio de Simulación de Eventos Discretos (SEDLab) pertenece al Instituto de Ciencias de la Computación (ICC UBA-CONICET) en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (FCEyN-UBA). Su objetivo es hacer avanzar el estado del arte en la construcción sistematizada de modelos de simulación, análisis y visualización de datos. Nuestro enfoque facilita y promueve el estudio interdisciplinario de sistemas complejos, que integran fenómenos naturales, sociales y computacionales.

https://modsimu.exp.dc.uba.ar/
@SEDLab_ICC
