Universidad del Desarrollo
Facultad de Ingeniería

DETECCIÓN DE TELETRABAJO UTILIZANDO DATOS XDR

¿Es el teletrabajo una tendencia creciente en Chile?

POR: LUIS ANDRÉS RAMÍREZ VERA

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Sra., Loreto Bravo, Sr. Leonardo Ferres

Diciembre 2022

SANTIAGO

Este trabajo está dedicado a mi esposa María José Gutiérrez, por su paciencia y motivación en cuanto a la realización de este Magister y durante todo el transcurso del mismo.

AGRADECIMIENTO

En primer lugar, le quiero agradecer a Dios por darnos salud y energía para llevar a cabo este proyecto. A mi familia por su constante motivación durante todo el transcurso del programa: mi esposa María José, mis hijos Joaquín y Alonso, mi suegra “Lelita”, mi mamá María, mi papá Luis Humberto Q.E.P.D., mis hermanos.

Quiero agradeceré a los profesores guía Loreto Bravo y Leo Ferres por sus orientaciones y valiosos feedbacks durante el transcurso del proyecto Capstone y a los profesores Alonso Astroza y Victor Navarro por su apoyo en las dificultades técnicas que surgieron. Adicionalmente le doy las gracias a todos los profesores del programa quienes aportaron el conocimiento necesario para llevar a cabo este proyecto.

También quiero agradeceré a Karin Becker, Isabel Alliende y Nicolás Duhart, por darme orientación y apoyo necesario por parte de la empresa para la realización de este Magister.

No quiero dejar de agradecer a muchos amigos y compañeros de trabajo que me motivan constantemente a dar lo mejor.

TABLA DE CONTENIDO

Resumen

Desde el año 2019, nuestro país ha estado en una situación bastante convulsionada debido a diversas movilizaciones sociales, donde hubo hechos que impactaron negativamente a la ciudadanía como es el caso de los daños ocurridos en el metro de Santiago que dejaron el 87% de la red sin servicio en octubre del 2019.

Tan solo 5 meses después, el día 18 de marzo del año 2020 se declaró la pandemia Covid-19, lo que trajo consigo cambios significativos para todo el planeta. Partiendo por los negativos efectos en la salud de las personas causando lamentables pérdidas humanas.

Uno de los cambio más relevantes fue la restricción prácticamente total del desplazamiento de las personas para evitar la propagación del virus que en aquellas fechas aun no tenía una cura disponible.

Ambos eventos causaron un gran impactado laboral donde empresas de diversos tamaños se vieron afectadas fuertemente, muchas pequeñas y medianas empresas tuvieron que cerrar sus puertas, y por su parte, muchas grandes empresas se vieron obligadas a reducir sus nóminas de trabajadores, debido a la incertidumbre del momento. De acuerdo a datos oficiales del Banco Central, en julio de 2020 hubo una tasa de desocupación del 13,1%, siendo la cifra más alta durante toda la pandemia. [1]

Debido a las restricciones de desplazamiento antes expuestas, las empresas tuvieron que realizar despliegues técnicos en tiempo record para continuar con las operaciones de negocio en modalidades de teletrabajo, algo que poco se había explorado hasta ese momento, pero que gracias a las tecnologías disponibles, se logró hacer frente en tiempos óptimos. Los planes de cambio de modalidad incluyeron capacitaciones para los trabajadores, implementación de herramientas de comunicación, trabajo a distancia y robustecimiento de las medidas de seguridad en las conexiones, entre otros.

Estos cambios trajeron consigo una disminución notoria del flujo de personas en transporte público y tránsito vehicular.

Con el paso del tiempo, se fueron eliminando las restricciones y las personas comenzaron a volver a sus rutinas diarias y en el mundo de las empresas, la modalidad de teletrabajo logró persistir incluso en momentos con mejores condiciones sanitarias.

En la actualidad, siendo la modalidad de teletrabajo una realidad más que una excepción, surgen nuevos campos de estudio que requieren ser analizados en torno a esta materia.

El presente informe, busca proponer un algoritmo de detección de dispositivos en modalidad de teletrabajo utilizando fuentes de datos XDR de una empresa de telecomunicaciones de Chile.

1. Introducción

Desde el año 2019 y hasta mediados del año 2022, nuestro país se vio afectado por eventos que afectaron el normal funcionamiento de las ciudades especialmente en temas de desplazamiento. Durante este periodo, una gran cantidad de empresas tuvieron que adoptar medidas extraordinarias y permitieron a sus trabajadores realizar teletrabajo debido a las estrictas restricciones sanitarias existentes en aquel momento.

Este año 2022, luego de 2 años y medio de pandemia y comenzando recientemente una nueva etapa, donde se han eliminado las mascarillas y los aforos, surge la interrogante de como esto impactará en el retorno presencial al trabajo y en la permanencia del teletrabajo.

Considerando lo expuesto anteriormente, no es fácil pronosticar como será la aplicación de normas de teletrabajo en cada empresa de cara al futuro, pero sí está claro que las condiciones no son las mismas pre y post pandemia, lo que inevitablemente puede traer ciertas consecuencias negativas, principalmente a los trabajadores [2], como por ejemplo el retorno, en ciertos casos, a los largos traslados casa-trabajo que durante la pandemia fueron evitados.

El teletrabajo antes de la pandemia era un tema poco explorado en Chile, a pesar de que existía bastante experiencia en el mundo, por ejemplo en Estados Unidos, Asia o Europa donde es común encontrar empresas que prestan servicios de profesionales a otros países (offshoring). Nuestras leyes no estaban adaptadas para atender las relaciones laborales del teletrabajo y con la llegada de la pandemia, todas las conversaciones que se encontraban en curso en el gobierno tuvieron que ser aceleradas, logrando en abril de 2020 promulgarse la Ley 21.220 que modifica el código del trabajo, incorporando regulaciones necesarias para el trabajo a distancia y el teletrabajo, destacándose entre sus características la seguridad de los trabajadores en el teletrabajo y el derecho a la desconexión digital.

El tema del teletrabajo no es algo nuevo a nivel mundial, de hecho el concepto de teletrabajo fue planteado por Margrethe Olson en 1983, en su paper “Remote office work: Changing work patterns in space and time” [3], donde identifica el teletrabajo como una de las opciones viables para considerando las alternativas de trabajar entre 1 día y una semana completamente desde remoto. Se destaca en este paper que hay ciertas profesiones con mayores posibilidades de realizar esta modalidad de trabajo. Esto indica que las profesiones de las personas influye en las posibilidades del teletrabajo.

Table I. The Research Sample. People Who Work-at-Home.

	Number
Clerical	
Data Entry Clerks	4
Professional	
Software Engineers/Programmers	7
Course Development Analysts	5
Loss Control Consultants	6
Staff Interviewers	2
Managerial	
Technical Managers	2
Staff Managers	
Project Managers	4

M. Olson, 1983

Este aspecto continua siendo analizado con el paso del tiempo, por ejemplo, un estudio realizado por Jonathan I. Dingel y Brent Neiman, denominado How many jobs can be done at home? [4] se analizaron las respuestas de un programa del Departamento del Trabajo, identificando que un 37% de los trabajos podría se realizado de forma remota.

En otro estudio realizado por la consultora McKinsey, revela que más de un 20% de la fuerza laboral podría trabajar de forma remota entre 3 a 5 días de manera eficiente como si estuvieran en una oficina. Si esto fuera aplicable, eso significa que habría 3 o 4 veces más personas teletrabajando que antes de la pandemia y esto tendría un profundo impacto en la economía urbana, transporte, gastos de consumo, entre otros.

En contraste, un 61% de la fuerza laboral, solo podría realizar laborales de forma remota solo algunas horas a la semana.

Adicionalmente McKinsey resalta que la viabilidad de teletrabajo no depende del cargo de los trabajadores sino más bien de las tareas y ocupaciones de cada uno de ellos. [7]

En este sentido surge una nueva interrogante: ¿Podrían las tendencias hacia el teletrabajo, impulsadas por los nuevos trabajadores y consumidores, incidir en la forma en que se desarrollan los trabajos? En otras palabras, las demandas de nuevos trabajadores respecto a contar con opciones de teletrabajo , ¿podría implicar que las empresas tengan que adaptarse a estas nuevas formas de trabajar?

Para investigar un poco más a fondo sobre este tema, el presente proyecto tiene por fin presentar la evolución del teletrabajo mediante el uso de datos de telefonía móvil anonimizada, y detectar si post-pandemia, los niveles de teletrabajo se han mantenido en el tiempo. Este estudio presente proyecto se enfocará en la región Metropolitana, que representa la mayor parte de habitantes del país. De acuerdo al último Censo realizado por INE en el año 2017, la Región Metropolitana tiene 8.310.964 habitantes lo que representa el 42% del total del país. [8]

Las conclusiones obtenida a partir de este análisis, podrían ser de gran utilidad para conocer hacia donde van las tendencias del teletrabajo con información concreta y visualizar sus posibles impactos en la ciudad.

Considerando que eventualmente el teletrabajo se puede comenzar a expandir como una de las opciones oficiales de trabajo en Chile, es importante tener mecanismos efectivos que permitan detectar y luego proyectar el crecimiento de este modo de trabajo.

Adaptándose al teletrabajo

2. Trabajos Relacionados

El teletrabajo ha sido un tópico en aumento en los últimos años en nuestro país y principalmente postpandemia.

De acuerdo al Instituto Nacional de Estadísticas INE en su Boletín complementario N°2 de *REMUNERACIONES Y COSTO DE LA MANO DE OBRA* [11] señala que en junio de 2020 un **28,9%** de los trabajadores realizó algún tipo de modalidad de teletrabajo en Chile, produciéndose una baja en diciembre de 2021 llegando a un **10,9%** de acuerdo a la versión N°9 correspondiente a la última versión de este boletín [12].

El inconveniente del apartado teletrabajo de este estudio, era la colección manual de datos mediante encuestas llenadas por las empresas, las que eran consolidadas, analizadas y publicas en los reportes respectivos. El problema común de las encuestas es que produce data con poca frecuencia y tiene un alto costo el llevarlas a cabo.

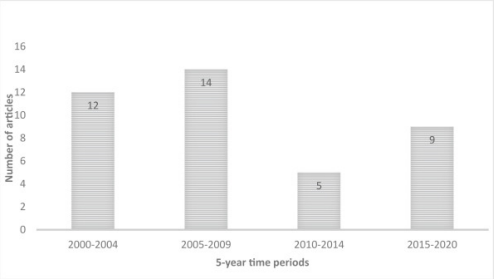
Considerando un análisis más automático y que considere otras fuentes de información, encontramos el estudio realizado por Yingkun, Chen, Junfan y Ming [13], quienes elaboraron un estudio de detección de casa y trabajo utilizando datos anonimizados de la red de celular CSD (Cellular Signalling Data) y realizaron la validación contra datos de voluntarios reales. Acá señalan la importancia de identificar las diversas variantes en cuanto horario laboral, dada la dificultad de conocer escenario de trabajadores nocturnos. Otro tema considerado es la identificación de 2 bloques horarios: horas de trabajo y horas de descanso, siendo estas mayoritariamente 10:00–16:00 y 20:00–6:00 respectivamente. Finalmente consideran la incorporación de Puntos de Interés (POI) dentro del algoritmo para permitir la detección más precisa de lugar (trabajo o casa), utilizando elementos de la ciudad como edificios, centros comerciales, etc.

Otro estudio interesante relacionado a la detección de domicilio mediante data de antenas de celulares, es el realizado por Pappalardo, Ferres, Sacasa, Cattuto y Bravo [9], donde también se realiza el contraste con data de personas reales y se realiza la comparación entre distintos tipos de datos de celulares (CDR, CPR y XDR) y se observa que XDR entrega resultados más precisos y estables, incluyendo las pruebas de minimización para determinar precisión con menor cantidad de datos.

La data de antenas de celulares ha demostrado ser una forma eficaz para la detección de diversos temas relativos a la ciudad, como los mencionados anteriormente y otros, como el estudio sobre atracción de las personas a centros comerciales y mixtura social realizado por Beiró, Bravo, Caro, Cattuto, Ferres y Graells-Garrido [15].

Dado que hasta ahora la suposición (assumption) estándar era la de considerar a la ubicación de la persona en lugares distintos entre trabajo y casa, es que aun no se han realizado muchos estudios en profundidad respecto al teletrabajo, más allá de considerar estudios cualitativos de diversa índole o encuestas dirigidas a las empresas o los trabajadores.

Esto se puede ver reflejado en es estudio realizado por Athanasiadou y Georgios [14], quienes analizaron la literatura relacionada al teletrabajo durante el transcurso de los últimos 20 años y presentaron las metodologías utilizadas para la recolección de los datos. El mecanismo más utilizado de recopilar datos respecto a personas teletrabajando se basa en encuestas.



Periodo de investigación

Research design	
Qualitative	12 (30%)
Quantitative	21 (52,5%)
Mixed	5 (12,5%)
Other	2 (5%)
Study methodology	
Case study	3
Experiment	2
Literature review	2
Survey	22
Interviews	9
Other	6

Diseño y metodología

Gracias a los antecedentes proporcionados por estos trabajos relacionados, podemos avanzar en la búsqueda de un algoritmo de detección del teletrabajo mediante data anonimizada de celulares, dada la confiabilidad de tiene este mecanismo, sin omitir aspectos tan relevantes como son la definición de bloques horarios, el uso de XDR asociado con información geoespacial de la ciudad y los procesos de minimización de data.

3. Hipótesis y Objetivos

Considerando que la pandemia abrió nuevas posibilidades relacionadas con el teletrabajo en nuestro país, ha aumentado el interés tanto a nivel de gobierno, empresas e investigación, por contar con herramientas apropiadas que permitan visualizar como se está produciendo este movimiento en las formas de trabajar.

Las hipótesis planteadas son 2:

1) El teletrabajo es una tendencia creciente y ha logrado mantenerse en el tiempo desde los inicios de la pandemia.

2) Los lugares donde se concentra el teletrabajo son las zonas más lejanas al centro de la ciudad.

Objetivos generales:

Analizar los datos anonimizados de movilidad XDR con el fin de detectar conexiones de dispositivos en modalidad de teletrabajo.

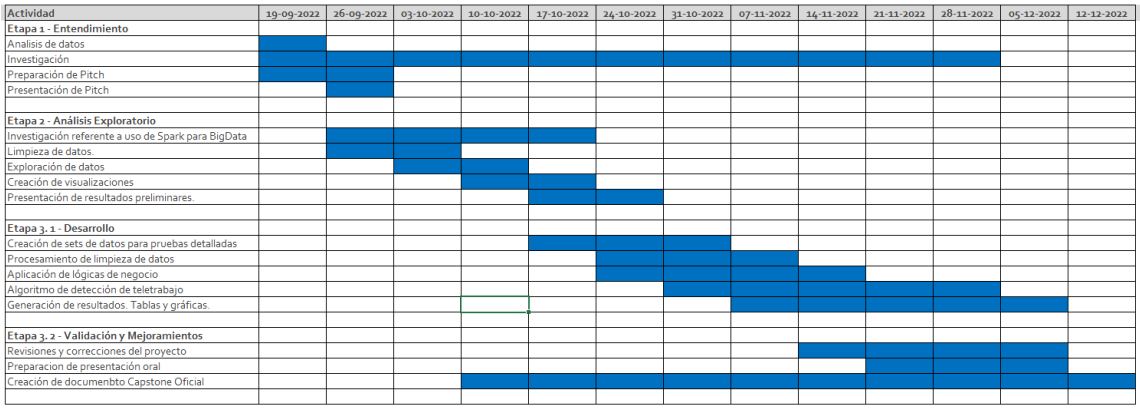
Algunos puntos de interés para este análisis podrían incluir la identificación la cantidad de teletrabajo durante el periodo en estudio (2020-2022), zonas con mayor y menos teletrabajo, antenas con mayor sobrecarga debido al teletrabajo, entre otros.

Estar atentos a los cambios mediante herramientas analíticas puede ayudar a adaptarse con antelación a nuevos escenarios no vistos hasta el momento.

Objetivos específicos:

- Analizar los datos XDR de entrada y prepararlos para su procesamiento.
- Establecer un proceso con configuraciones óptimas de servidor para el procesamiento de altos volúmenes de datos.
- Creación de un algoritmo que permita detectar si un dispositivo se encuentra o no en teletrabajo.
- Identificar en que cantidad de datos se logran los mejores resultados de detección de teletrabajo.
- Generar resultados que permitan analizar de forma numérica y gráfica, los lugares de la región metropolitana que mantienen las modalidades de teletrabajo a lo largo del tiempo.

A continuación se presenta el plan de trabajo desarrollado para llevar a cabo este proyecto:



Actividad	19-09-2022	26-09-2022	03-10-2022	10-10-2022	17-10-2022	24-10-2022	31-10-2022	07-11-2022	14-11-2022	21-11-2022	28-11-2022	05-12-2022	12-12-2022
Etapa 1 - Entendimiento													
Analisis de datos													
Investigación													
Preparación de Pitch													
Presentación de Pitch													
Etapa 2 - Análisis Exploratorio													
Investigación referente a uso de Spark para BigData													
Limpieza de datos.													
Exploración de datos													
Creación de visualizaciones													
Presentación de resultados preliminares.													
Etapa 3. 1 - Desarrollo													
Creación de sets de datos para pruebas detalladas													
Procesamiento de limpieza de datos													
Aplicación de lógicas de negocio													
Algoritmo de detección de teletrabajo													
Generación de resultados. Tablas y gráficas.													
Etapa 3. 2 - Validación y Mejoramientos													
Revisiones y correcciones del proyecto													
Preparacion de presentación oral													
Creación de documenbto Capstone Oficial													

Plan de trabajo - Carta Gantt

4. Datos y Metodología

4.1. Datos

La fuente de información para este proyecto corresponde a datos anonimizados de telefonía del tipo XDR (eXtended Detail Records) que contienen los registros de transmisión de datos entre dispositivos y antenas telefónicas.

La estructura de la información consta de un identificador **anónimo de dispositivo, el timestamp de ejecución y las coordenadas de latitud y longitud del evento**.

device	datetime	lon	lat
7b6a028c9394f4de8...	2021-03-04 13:07:14	-70.52948	-33.38543
290425bbe367b612d...	2022-03-03 08:18:27	-71.54372	-32.92978
c0d3aad4423908178...	2020-03-05 05:06:00	-71.4891	-32.72972
c20f978a8d267567c...	2021-03-03 04:33:39	-70.72083	-32.74917
88d1867b0e40e20c0...	2022-03-03 11:28:36	-70.57401	-33.58549
342839e67b2da7570...	2022-03-01 03:50:06	-71.38656999999999	-34.17699
79b1463aa16aa1dd9...	2022-03-02 23:03:51	-71.69481999999999	-33.398990000000005
74102fb7c6a5797e1...	2021-03-04 12:02:23	-71.19716	-30.58822
5511354de5fbe3ed4...	2022-03-01 18:07:50	-71.21537	-33.684740000000005
72a202eb4ce05ea3e...	2022-03-01 01:18:32	-70.5951	-33.4248

Estructura de un archivo XDR

La data se disponibilizó en 499 archivos de tipo parquet los cuales contienen en total ***3.465 millones de registros***.

Los datos disponibles corresponden a 12 días de conexiones XDR del mes de marzo, distribuidos entre los años 2020, 2021 y 2022.

Debido a la gran cantidad de datos a ser analizados, es necesario realizar el procesamiento en un servidor de gran capacidad y usar la herramienta Spark. Para esto, se utilizó el servidor Centella disponibilizado por la UDD.

También se utilizan datos de información geográfica para poder posicionar los resultados identificados. En este caso se están utilizando archivos shape de las comunas del país, obtenidas desde el sitio de la biblioteca del congreso nacional. [10]

4.2. Metodología

La metodología aplicada consiste en realizar análisis cuantitativos con fuentes de datos XDR generados de manera automática, ya sea por el usuario al acceder a servicios de internet de manera intencional o por las aplicaciones instaladas en el dispositivo que de forma automática se conectan frecuentemente a los servicios de internet.

Dentro de los alcances se considera la identificación de dispositivos que se encuentren en modalidad de teletrabajo, comunas con mayor teletrabajo y antenas con mayor presencia de dispositivos en teletrabajo.

Por temas de volumen de datos, este proyecto se enfocará en tratar de **identificar modalidad de teletrabajo el primer día miércoles de los años 2020, 2021 y 2022 en la región metropolitana**.

A continuación se muestra un diagrama con el proceso realizado.

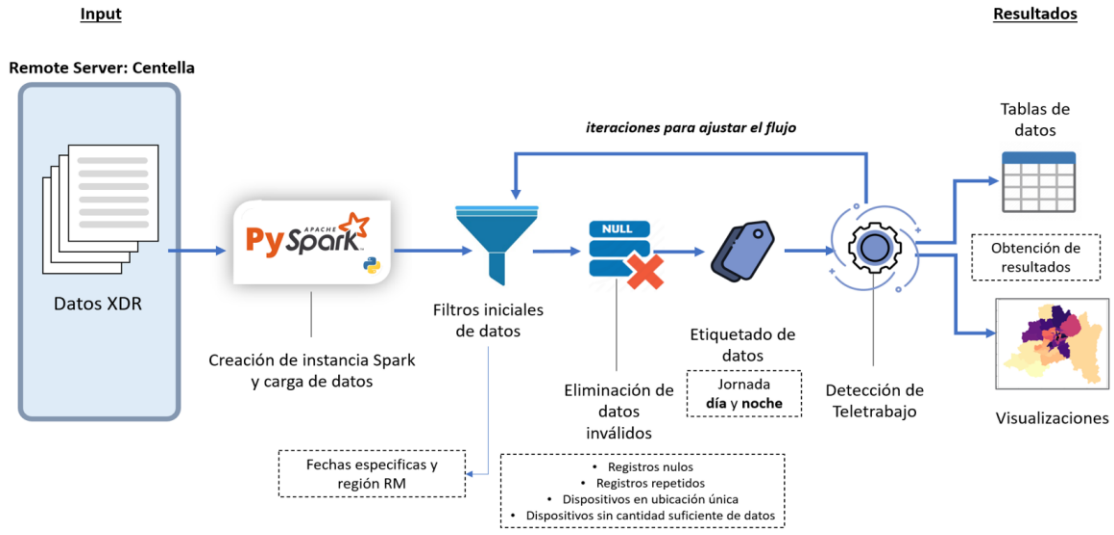


Diagrama del proceso de detección de teletrabajo

Este proceso consiste en las siguientes etapas:

1. **Proceso de carga de datos:** Generación dinámica de los nombres de archivo a ser importados.

   Para ir testeando distintos escenarios de carga de archivos, se creó una variable “cantidad_archivos” donde se puede especificar la cantidad de archivos a procesar.

i. Por una parte, el testeo con una baja cantidad de archivos permite mayor rapidez de procesamiento pero se cuenta con poca cantidad de datos para obtener un análisis adecuado.
ii. Por otra parte, el testeo con una alta cantidad de archivos permite mayor precisión de datos pero puede requerir varias horas de procesamiento.
iii. Por esta razón es importante elegir un numero óptimo de archivos, para ir desarrollando las pruebas.

Nota: Los archivos fuente tiene una estructura fija, considerando un correlativo entre 1 y 499 al final del nombre del archivo.

2. **Filtro inicial de datos:** En esta etapa se realiza el filtrado de los datos originales para disponer únicamente de los datos que serán utilizados en el proceso.

    a. Para efectos de este proyecto se filtran solo los datos de la Región Metropolitana.
    b. Para medir la evolución año tras año, se considerará el análisis del primer miércoles de marzo de los años 2020, 2021 y 2022.

3. **Eliminación de datos inválidos:**

    a. Se procede a eliminar los datos nulos o inválidos. (Limpieza de datos)
    b. Se eliminan los datos repetidos.

4. **Etiquetado de datos:**

    a. En esta etapa y teniendo los datos depurados, se realiza el etiquetado de los datos indicando la jornada a la que pertenecen, ya sea día o noche, de acuerdo a un horario definido.

5. **Detección de teletrabajo:**

    a. En esta etapa se importan datos espaciales que sean de relevancia para realizar el análisis. En este caso, se consideró la data geo espacial de las comunas del país, obtenida desde el sitio web de la biblioteca del congreso nacional de Chile.
    b. A continuación, se realizan algunas operaciones de join Geo espacial para asociar los datos de entrada con las comunas de la región metropolitana.

c. Se realizan los algoritmos de calculo para excluir datos que se han definido como inválidos desde el punto de vista de reglas de negocio:
    i. Dispositivos con baja cantidad de registros.
    ii. Dispositivos con una única ubicación.
d. Se aplica la lógica de identificación del teletrabajo.
e. Finalmente se obtienen los primeros resultados en formato de tablas y visualizaciones.

En este punto se revisa la coherencia de los resultados y luego se ajusta el programa en caso de ser necesario.

6. **Obtención de resultados:** Al obtener un resultado óptimo y de acuerdo a lo esperado, y procede a testear con distinta cantidad de datos.

   El proceso ha sido ejecutado sobre una base acotada de registros considerando el procesamiento de 50, 100, 200 y 499 archivos.

   Cabe mencionar, que con la base de datos completa, se requiere una gran cantidad de tiempo de procesamiento.

   Los procesamientos se realizaron considerando un día informado como parámetro, y el proceso se repite por los días que se necesiten evaluar.

A continuación se mencionan diversos aspectos considerados en la metodología que permiten preparar los datos para llegar a la detección del teletrabajo.

Mejorar la estructura de información

Con el fin de realizar un análisis más detallado de la información relacionada al tiempo, se agregaron nuevas columnas derivadas que aportan valor al análisis.

En base a la columna ***timestamp***, se incorporaron las siguientes columnas:

- date: Sólo fecha
- time: Sólo hora
- day: Número de día
- month: Número de mes
- year: Año
- day_text: Nombre abreviado del día de la semana
- day_of_week: Número de día de la semana

- hour: Hora
- minute: Minuto
- second: Segundo

date	time	day	month	year	day_text	day_of_week	hour	minute	second
2021-03-04	13:07:14	4	3	2021	Thu	5	13	7	14
2022-03-03	08:18:27	3	3	2022	Thu	5	8	18	27
2020-03-05	05:06:00	5	3	2020	Thu	5	5	6	0
2021-03-03	04:33:39	3	3	2021	Wed	4	4	33	39
2022-03-03	11:28:36	3	3	2022	Thu	5	11	28	36
2022-03-01	03:50:06	1	3	2022	Tue	3	3	50	6
2022-03-02	23:03:51	2	3	2022	Wed	4	23	3	51

Estructura de datos con adaptaciones

Luego de varias iteraciones de revisión, se opta por mantener únicamente las siguientes columnas que serán utilizadas durante el algoritmo de detección de teletrabajo. Esto para optimizar el data set que será procesado.

- day: Número de día
- day_of_week: Número de día de la semana
- hour: Hora

Definición de los bloques horarios día y noche

Para poder identificar dispositivos en teletrabajo, se agrega una columna que indica el ***bloque horario*** de cada evento/registro. Los bloque definidos son: día y noche.

A diferencia de otros modelos de detección de ubicación usando data de celular [13], donde se excluyen ciertos horarios de traslado dada la suposición de que existe distinta ubicación entre casa y trabajo, para este modelo esta exclusión no se aplica y se mantienen los límites exactos entre cada bloque. En este caso se considera el supuesto de que la persona en teletrabajo se mantiene en el mismo lugar. Por otra parte, el movimiento por traslado es algo natural en el trabajo presencial y que aporta valor a la diferenciación.

Los bloques horarios definidos son los siguientes:

- Día: 7:00 a 19:00
- Noche: 19:01 a 6:59

Importante: La evaluación de estos bloques horarios solo se debe realizar durante días hábiles. Por lo anterior se deben excluir feriados y fines de semana. Este aspecto es excluido de este análisis debido a que se cuenta con cantidad acotada de días. Pero en la aplicación real, esta validación debe ser incorporada.

Separación de datos para el análisis

Considerando que el data set original tiene las coordenadas de los dispositivos en términos de latitud y longitud, pero no tienen la región o comuna identificada, es necesario realizar una unión geoespacial de la información.

Para realizarlo, se obtuvo el archivo shape file de las comunas del país [10], el cual cuenta identificación de comunas y regiones del país. Para efectos de este proyecto, solo se consideran las comunas de la Región Metropolitana.

Ambos set de datos se levantan como GeoDataFrames y se realiza un Join Geoespacial para poder filtrar los datos XDR que solo correspondan a la Región Metropolitana.

Luego se continua trabajando con los datos resultantes de esta unión.

Exclusión de dispositivos sin movimiento

Uno de los casos que debe ser analizado en mayor profundidad es el caso de dispositivos que se encuentren 100% en una misma ubicación. Para efectos de este proyecto, se toma el supuesto de que el dispositivo debe tener al menos 1 movimiento entre distintas antenas para para ser un elemento valido, tratando de reflejar un escenario de la vida real dado que una persona naturalmente debería tener movimientos fuera de su domicilio.

Considerando que analizaremos los primeros miércoles de los años 2020, 2021 y 2022, se investigó sobre la situación país en aquellos días:

Marzo 2020: El país se encontraba sin pandemia pero aún permanecían los problemas de desplazamiento debido a estaciones de metro fuera de servicio. El día martes 3 de marzo se detectó el primer caso de Covid-19 en Chile y el día 18 de marzo se declaró estado de emergencia. Antes de esta fecha no existía restricción de movilidad total.

Marzo 2021: El país venia regresando del periodo de vacaciones y se encontraban todas las comunas de la RM en fase 3 o 4 (estado sin Cuarentena), con posibilidades de desplazamiento. El día jueves 6 de marzo de 2021 se ajustaron las medidas debido al aumento de casos y se movieron gran parte de las comunas a fase 2.

Marzo 2022: El país se encuentra en proceso de vacunación de la 4ta dosis y sin restricciones de movilidad aplicadas.

Dado esto, no es aplicable el caso de que por motivos de cuarentena o restricción total, las personas no podían salir de sus casas.

Por esta razón, el algoritmo de detección analiza todos los dispositivos que no tienen movimiento y los excluye del análisis. Se asume para estos casos, que los dispositivos corresponden a conectividad inmóvil y que no son dispositivos de personas. Algunos de estos dispositivos pueden ser: cámaras, dispositivos smart home, señales de banda ancha móvil y otros similares.

Algoritmo de detección

Una vez preparados los datos, el método de detección identifica la cantidad de conexiones por antena por cada dispositivo en sus respectivos bloques horarios. La cantidad de conexiones por antena puede variar de dispositivo en dispositivo. Para descartar dispositivos con baja conectividad se consideran solamente aquellos dispositivos que tienen al menos **10 o más** conexiones a la misma antena. Esto asegura una conectividad en un lugar de al menos 3 horas. Al analizar los registros XDR se lograr detectar que la generación se realiza cada 20 o 30 minutos por cada dispositivo.

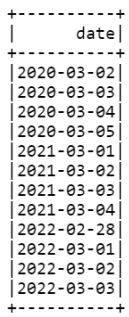
Una vez identificados los puntos de conexión más frecuentes en el día y la noche, se comparan entre sí y en caso de ser iguales, se marca el dispositivo como “teletrabajo”. Esto se realiza de manera automática con todos los dispositivos a la vez.

5. Resultados

Análisis exploratorio

En la primera etapa del proceso, revisamos los datos con los que contamos.

Las fechas disponibles en el data set de origen son las siguientes:



```
+----------+
|      date|
+----------+
|2020-03-02|
|2020-03-03|
|2020-03-04|
|2020-03-05|
|2021-03-01|
|2021-03-02|
|2021-03-03|
|2021-03-04|
|2022-02-28|
|2022-03-01|
|2022-03-02|
|2022-03-03|
+----------+
```

Años disponibles en los datos de origen

Esta fechas corresponden a la primera semana de marzo considerando días de lunes a jueves en todo el país.

La cantidad de datos por día disponibles en el data set completo son los siguientes:

Día	Fecha	Cantidad de registros
Lunes	2020-03-02	13.877.907
Martes	2020-03-03	146.080.117
Miércoles	**2020-03-04**	**147.246.250**
Jueves	2020-03-05	133.982.067
Lunes	2021-03-01	15.551.783
Martes	2021-03-02	163.725.155
Miércoles	**2021-03-03**	**165.347.303**
Jueves	2021-03-04	149.855.748
Lunes	2022-02-28	79.296.163
Martes	2022-03-01	823.278.220
Miércoles	**2022-03-02**	**850.588.520**

Jueves	2022-03-03	776.883.215
Cantidad total de registros:		**3.465.712.448**

Tabla resumen con la cantidad de datos por día

Acá podemos observar 2 situaciones que se dan con los datos:

1) Los días lunes de cada año, solo contienen una porción aproximada del **9%** en comparación con los otros días entregados.

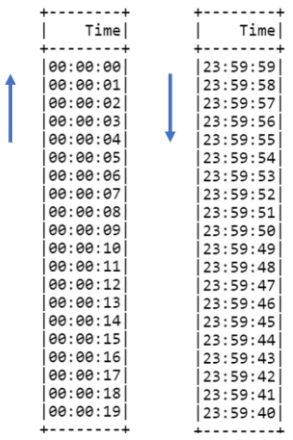
2) Los datos de los años 2020 y 2021, cuentan con menos cantidad de datos que el año 2022.

Esta disparidad en los datos podría afectar la identificación de dispositivos en teletrabajo, por lo que revisaremos los resultados en detalle más adelante.

Revisión del campo tiempo

Dado que la ***hora*** es un dato relevante para el análisis, aseguramos que se encuentren disponibles datos en todas las horas del día. El resultado es correcto y tenemos datos desde las 0 a las 23 hrs.

Adicionalmente validamos que los datos estén distribuidos con una frecuencia adecuada. En este caso podemos ver que hay datos reportados a cada segundo al inicio y fin de cada día.



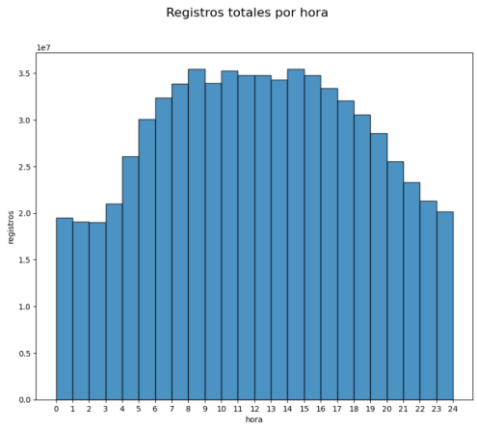
Time
00:00:00
00:00:01
00:00:02
00:00:03
00:00:04
00:00:05
00:00:06
00:00:07
00:00:08
00:00:09
00:00:10
00:00:11
00:00:12
00:00:13
00:00:14
00:00:15
00:00:16
00:00:17
00:00:18
00:00:19

Time
23:59:59
23:59:58
23:59:57
23:59:56
23:59:55
23:59:54
23:59:53
23:59:52
23:59:51
23:59:50
23:59:49
23:59:48
23:59:47
23:59:46
23:59:45
23:59:44
23:59:43
23:59:42
23:59:41
23:59:40

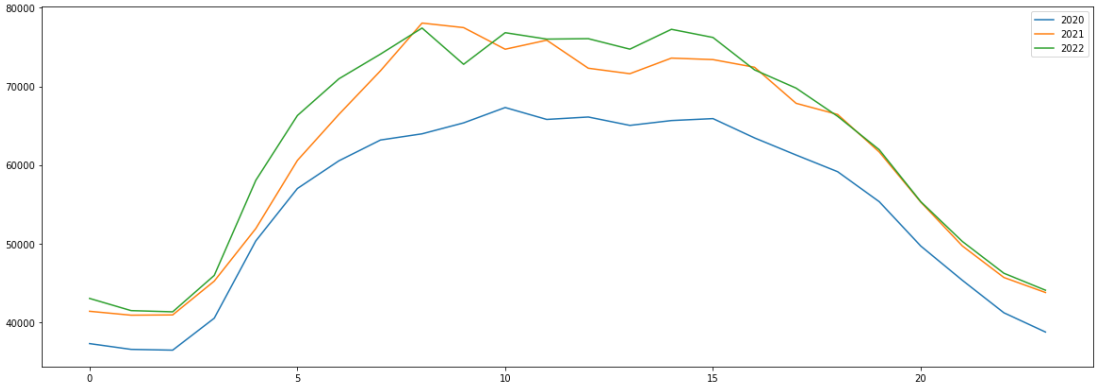
Verificación de horas en los datos de origen

Ejecución del proceso con archivos parquet

La cantidad de datos y dispositivos utilizados para ejecutar los procesos variará dependiendo de la cantidad de archivos cargados. Las pruebas se realizaron generalmente con 5, 50, 100 y 200 archivos parquet. **La ejecución final, una vez completado el proceso de desarrollo se realiza sobre la totalidad de los registros, es decir 499 archivos**.



Cantidad de datos por hora – Todos los años



Análisis de conexiones por hora de día por cada año

En este gráfico se puede apreciar una concordancia en las horas de conexión de todos los años.

Análisis de movilidad de dispositivos

Uno de los aspectos a evaluar en este análisis es identificar los dispositivos móviles e inmóviles,

Esto para poder identificar los casos en que los dispositivos son utilizados por personas o por otros dispositivos inmóviles, ya sea para empresas, dispositivos smart home u otros.

En este caso se realiza el siguiente análisis que consta de 2 etapas:

1) Primero se identifica la cantidad de ubicaciones distintas por cada dispositivo. Esto nos entrega una tabla con el siguiente formato:

device	lon	lat	count
5f7483349d2dec67c...	-70.52948	-33.38543	3
13d07cfb0edf933e4...	-71.54372	-32.92978	4
502bbd403ab68b694...	-71.4891	-32.72972	23
18e114bcad3458831...	-70.72083	-32.74917	2
d3ae2c76d643efa0b...	-70.57401	-33.58549	5
f1d83ef17cd7343d6...	-71.38656999999999	-34.17699	3
8e78099fce8c7433c...	-71.69481999999999	-33.398990000000005	9
5a53ba6e990e30380...	-71.19716	-30.58822	5
f09adbb7812de7c29...	-71.21537	-33.684740000000005	1
b9290be1aaba69273...	-70.5951	-33.4248	3

Análisis de ubicaciones por dispositivo

La columna **count** indica la cantidad de conexiones en un punto especifico por cada dispositivo. En caso de que la cantidad sea menor a 10, los dispositivos son excluidos del análisis.

2) Luego se calcula la cantidad de ubicaciones/antenas distintas por cada dispositivo.

device	antenas
6e37d69c1edfad403...	1
b6f29e4548a75aa83...	2
bedfb7114522e2efd...	2
6fe62f0a215f188f0...	1
0fef8d90d0bc49190...	1
24c95fe95673a9d88...	4
b120a8eb9b2cea8fd...	1
ee027f48ab82fc169...	1
9b41d7503a81cd3f9...	3
97095aa3b15a1ebea...	2

Análisis de antenas por dispositivo

Como resultado se ha aplicado el supuesto de que un dispositivo que tiene 1 ubicación detectada en la etapa 2 de este proceso, es identificable como un dispositivo inmóvil, por lo que es excluido del análisis completo.

Información Geográfica

Para poder desplegar información geoespacial correspondiente a ubicaciones, se utilizó Geopandas para la incorporación visual de mapas y coordenadas de interés. Desde el sitio web de la Biblioteca del Congreso Nacional de Chile (BCN) se obtuvieron los archivos shape de las comunas del país y la red vial. [10]

Presentación de resultados

Tras ejecutar el proceso de detección de dispositivos en teletrabajo con la **base de datos completa** para los primeros miércoles de los años 2020, 2021 y 2022, podemos apreciar los primeros resultados.

DETECCIÓN DE TELETRABAJO

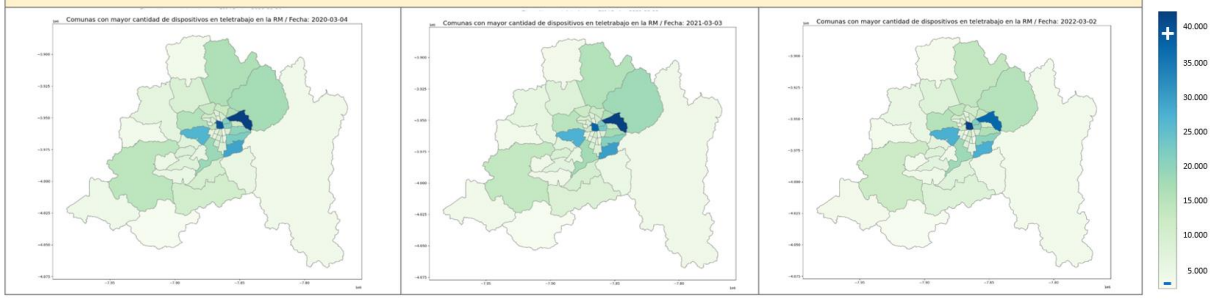
2020	2021	2022
1er Miércoles de Marzo en la RM		
		
Dispositivos en teletrabajo: **246.692**	Dispositivos en teletrabajo: **309.872**	Dispositivos en teletrabajo: **505.084**

Tabla de resultados con identificación de dispositivos en teletrabajo

En esta gráfica se aprecia a simple vista que el teletrabajo ha permanecido constante en el tiempo y se concentra en ciertas comunas específicas de la ciudad. Respecto a los datos

es posible apreciar que en el año 2022 se logró una mayor cantidad de identificaciones de teletrabajo.

Para entender un poco mejor los resultados, podemos revisar la siguiente tabla que muestra en términos numéricos los resultados de cada año.

Año	Cantidad de dispositivos únicos	Dispositivos en teletrabajo	% de teletrabajo detectado
2020	1.416.948	246.692	17%
2021	1.505.176	309.872	21%
2022	1.466.007	505.084	34%

Análisis cuantitativo de los resultados

Dispositivos en teletrabajo

El resultado de este proceso, también nos entrega los dispositivos identificados en teletrabajo, con lo cual se pueden realizar análisis en mayor profundidad para conocer el historial de movimiento de estos dispositivos y con ello poder determinar los ahorros de tiempo en desplazamiento, cuando se ejecuta el teletrabajo. Dado que contamos con la identificación del dispositivo, también es posible obtener las coordenadas de las antenas detectadas como teletrabajo.

```
+--------------------+
|              device|
+--------------------+
|019ef0909fceff505...|
|02ccc5888c3605db4...|
|1258832af43ddea7a...|
|12f3addecfe87a22e...|
|17385c63fbdc47b94...|
|1941d0597d823f257...|
|1ca4141b6d28ce5e2...|
|1f483e0f335580769...|
|1f679ac8239c33d0c...|
|21409f06bd28af318...|
+--------------------+
only showing top 10 rows
```

Tabla de dispositivos detectados en teletrabajo

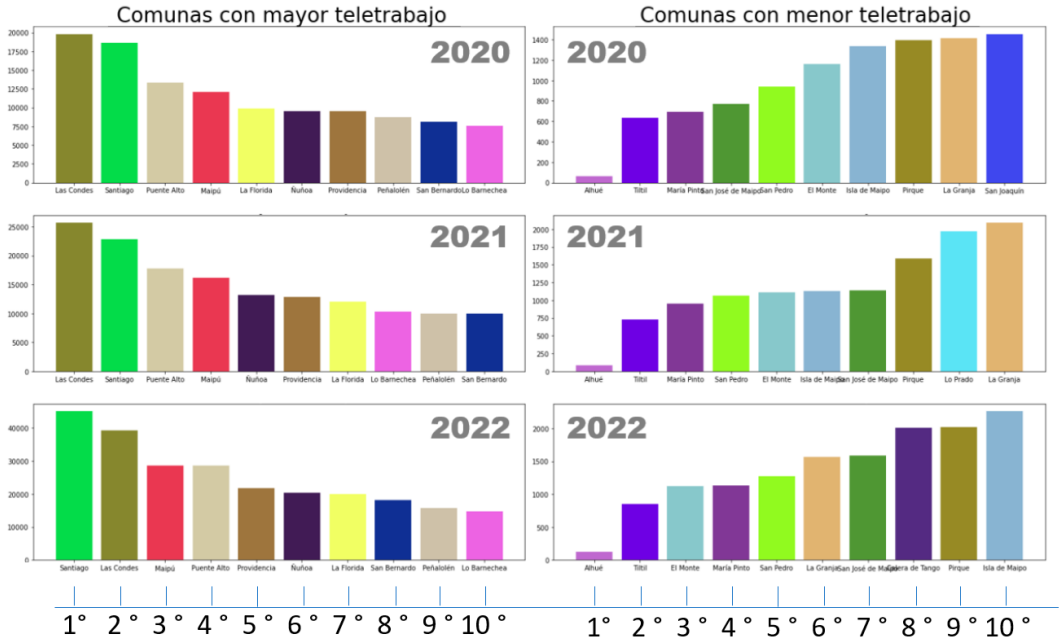
Ranking de Comunas en teletrabajo

Adicionalmente podemos obtener la cantidad de dispositivos en teletrabajo por comuna.

Comuna	dispositivos_en_teletrabajo
Santiago	44985
Las Condes	39246
Maipú	28620
Puente Alto	28620
Providencia	21762
Ñuñoa	20385
La Florida	19892
San Bernardo	18175
Peñalolén	15731
Lo Barnechea	14725
Pudahuel	14679
Quilicura	13202
Colina	12746
Vitacura	12537
Melipilla	11043
Estación Central	10657
Recoleta	10328

Ranking de teletrabajo por comuna – Año 2022

Podemos visualizar la evolución del teletrabajo por comuna año tras año en el siguiente gráfico de barras:



Análisis del teletrabajo por año y comuna

Es posible apreciar que las comunas que concentran mayoritariamente el teletrabajo son Santiago y Las Condes que son comunas consideradas céntricas y le siguen inmediatamente 2 comunas más lejanas como son Puente Alto y Maipú.

Indicadores del proceso

Finalmente y como complemento, el proceso va recopilando diversos indicadores que permiten entender el paso a paso del proceso de identificación, esto puede ayudar a entender el sentido de ciertos resultados y se pueden agregar nuevos indicadores que sean necesarios.

	indicador	valor
0	fecha_inicio	2022-11-28 06:48:40.789830
1	cantidad_archivos	50
2	filtro_fecha	2022-03-02
3	registros_origen	69199522
4	cantidad_antenas	4387
5	cantidad_antenas_rm	1577
6	datos_rm	28103761
7	dispositivos_distintos	3799267
8	conexiones_frecuentes_dia	302364
9	conexiones_frecuentes_noche	186117
10	informacion_teletrabajo	77315
11	dispositivos_en_teletrabajo	77315
12	fecha_fin	2022-11-28 07:12:49.468429
13	tiempo_de_duracion_del_proceso	0 days 00:24:08.678599

Ejemplo de captura de indicadores del proceso

6. Conclusiones

Como resultado, podemos observar que es factible realizar procesos de detección de dispositivos en teletrabajo de una manera eficiente y relativamente rápida mediante el uso de tecnologías modernas de Big Data. Los resultados expuestos presentan una consistencia a lo largo del tiempo y la evolución positiva hacia el año 2022 puede deberse a la madurez que van adquiriendo las empresas y los trabajadores en cuando a adoptar el teletrabajo como una forma viable de trabajo.

Podemos apreciar que en el año 2022, no existe una disminución del teletrabajo por lo que esta forma de trabajo podría tener proyección positiva hacia el futuro.

Una buena señal que se observa, es que el teletrabajo está siendo adoptado tanto en comunas centrales como en comunas mas alejadas del centro de la ciudad, lo que claramente vislumbra un buen horizonte en lo que respecta a la calidad de vida de los trabajadores que deben recorrer largas distancias para llegar a sus lugares de trabajo.

El actual proyecto tiene amplias oportunidades de mejora y considero que implementar nuevos mecanismos de detección del teletrabajo que sean eficientes y de bajo costo, es super relevante para poder medir la evolución de las ciudades en este aspecto que tiene un alto impacto laboral y también social.

Limitaciones del modelo

Una de las limitaciones del modelo actual es la cantidad de fechas consideradas en el análisis, lo que impide ver una evolución más amplia a lo largo del tiempo. Teniendo en cuenta que cada fecha contiene más de 100 millones de registros como base a nivel nacional, la ejecución del proceso requiere un considerable tiempo de procesamiento. Para ejecuciones completas, el tiempo requerido fue de entre 6 y 24 hrs. de procesamiento dependiendo de la disponibilidad del servidor. Por lo anterior, llevar a cabo esta ejecución en la práctica debe considerar algún tipo de optimización que puede ser aplicada mediante un pre-procesamiento que permita limpiar los datos en la medida que se van recopilando, a fin de llegar al procesamiento completo con datos depurados. Esto podría suponer un importante ahorro de tiempos.

Otro de los aspectos que no es posible distinguir, son las situaciones distintas al teletrabajo y que podrían tener comportamientos similares, por ejemplo:

- Colegio a distancia
- Universidad a distancia
- Estar sin trabajo

Entre otros.

Para poder avanzar en este aspecto, sería recomendable aplicar la metodología de contrastar los resultados con personas de prueba reales que cumplan los distintos roles: trabajadores, estudiantes, etc , fórmula que ha sido aplicada en otros modelos de detección previamente.

Trabajos futuros

Una mejora interesante a incorporar en futuras versiones sería la identificación del lugar del trabajo real, y de este modo sería posible medirá cual es el impacto positivo que tiene el teletrabajo en términos de tiempos de desplazamiento ahorrados.

Otro aspecto que se puede contemplar en un trabajo futuro, es incorporar los elementos de la ciudad como rutas, centros laborales, centros comerciales, u otros para mejorar la personalización de las zonas donde se realiza el teletrabajo.

Bibliografía

1. Banco Central, tasas de desempleo mensual https://si3.bcentral.cl/siete/ES/Siete/Cuadro/CAP_EMP_REM_DEM/MN_EMP_REM_DEM13/ED_TDNRM2?idSerie=F049.DES.TAS.INE9.10.M

2. Felstead, Alan & Jewson, Nick & Phizacklea, Annie & Walters, Sally. (2006). Opportunity to Work at Home in the Context of Work-Life Balance. Human Resource Management Journal. 12. 54 - 76. 10.1111/j.1748-8583.2002.tb00057.x.

3. Olson, Margrethe H. "Remote office work: Changing work patterns in space and time." Communications of the ACM 26.3 (1983): 182-187.

4. Dingel, Jonathan & Neiman, Brent. (2020). How many jobs can be done at home?. Journal of Public Economics. 189. 104235. 10.1016/j.jpubeco.2020.104235.

5. Bhattacharjee, Sujayita. (2020). 'Work from home' as an alternative to daily commuting for working women. Human Geographies. 14. 255-265. 10.5719/hgeo.2020.142.5.

6. Eliasson, Jonas & Mattsson, Lars-Göran. (2006). Equity Effects of Congestion Pricing: Quantitative Methodology and a Case Study for Stockholm. Transportation Research Part A: Policy and Practice. 40. 602-620. 10.1016/j.tra.2005.11.002.

7. McKinsey: What's next for remote work: An analysis of 2,000 tasks, 800 jobs, and nine countries https://www.mckinsey.com/featured-insights/future-of-work/whats-next-for-remote-work-an-analysis-of-2000-tasks-800-jobs-and-nine-countries

8. INE, Censo 2017 https://www.ine.cl/estadisticas/sociales/censos-de-poblacion-y-vivienda/censo-de-poblacion-y-vivienda

9. Pappalardo, L., Ferres, L., Sacasa, M. et al. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. EPJ Data Sci. 10, 29 (2021). https://doi.org/10.1140/epjds/s13688-021-00284-9

10. Mapas Vectoriales – Biblioteca del Congreso Nacional de Chile BCN. Obtención de archivos tipo shape de:"División comunal: polígonos de las comunas de Chile" y "Red vial: polilíneas de los caminos de Chile". https://www.bcn.cl/siit/mapas_vectoriales/index_html

11. INE - Boletín complementario N°2 de REMUNERACIONES Y COSTO DE LA MANO DE OBRA. https://www.ine.gob.cl/docs/default-source/sueldos-y-salarios/boletines/espa%C3%B1ol/base-anual-2016-100/m%C3%B3dulo-covid-19-ir-icmo/bolet%C3%ADn_covid_jjas.pdf

12. INE - Boletín complementario N°9 de REMUNERACIONES Y COSTO DE LA MANO DE OBRA. https://www.ine.gob.cl/docs/default-source/sueldos-y-salarios/boletines/espa%C3%B1ol/base-anual-2016-100/m%C3%B3dulo-covid-19-ir-icmo/bolet%C3%ADn-covid-19-(septiembre-a-diciembre-2021).pdf

13. Yingkun Yang, Chen Xiong, Junfan Zhuo, Ming Cai, "Detecting Home and Work Locations from Mobile Phone Cellular Signaling Data", Mobile Information Systems, vol. 2021, Article ID 5546329, 13 pages, 2021. https://doi.org/10.1155/2021/5546329

14. Athanasiadou, Chrisalena & Georgios, Theriou. (2021). Telework: Systematic Literature Review and Future Research Agenda. Heliyon. 7. e08165. 10.1016/j.heliyon.2021.e08165.

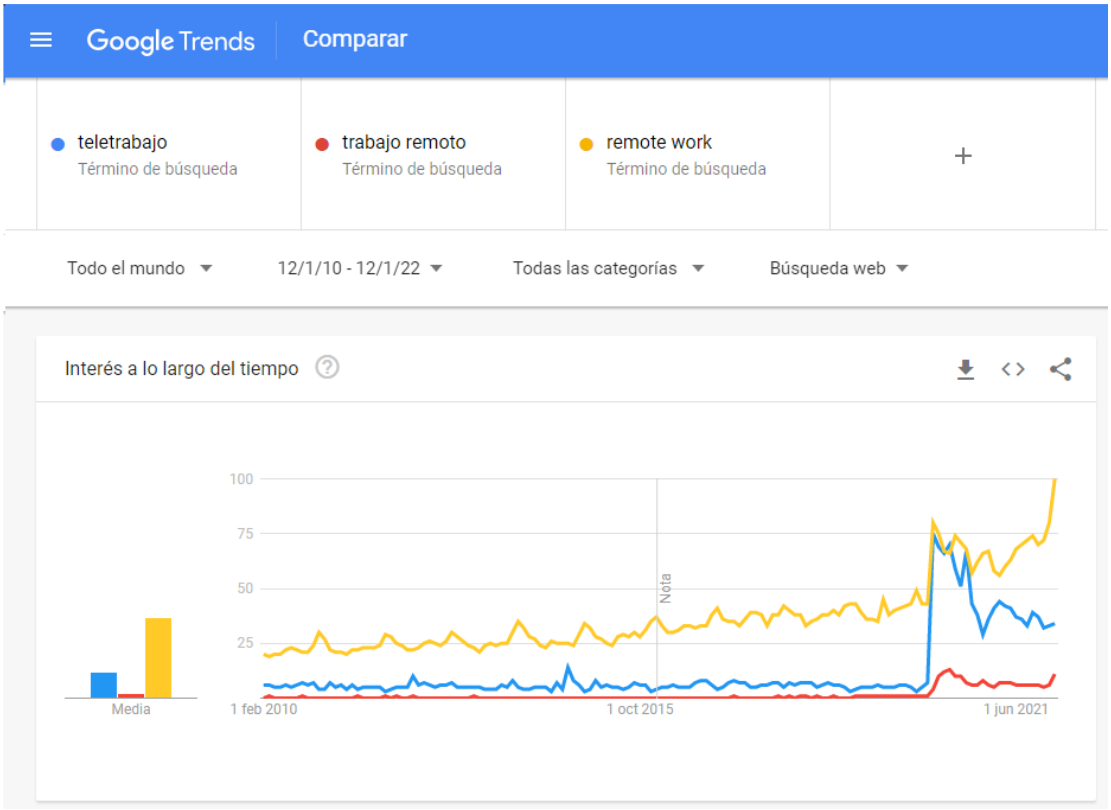
15. Beiró, M.G., Bravo, L., Caro, D. et al. Shopping mall attraction and social mixing at a city scale. EPJ Data Sci. 7, 28 (2018). https://doi.org/10.1140/epjds/s13688-018-0157-5

Anexos:

Como material complementario, se incluye el análisis de tendencias de búsqueda Google de teletrabajo, trabajo remoto o remote work en el mundo.

En este análisis se puede apreciar que ya existía una tendencia constante de búsquedas de teletrabajo durante los últimos 10 años en el mundo, excepto en América Latina (sin considerar Brasil) y Africa. Sin embargo, las búsquedas de “teletrabajo” en Latinoamérica se inician fuertemente recién en enero 2020 y post-pandemia se mantiene un 20% superior a números pre-pandemia. Dado esto, se podría pensar que es un fenómeno relativamente nuevo en este lado del continente.

Gráfica 1: Comparación de términos relacionados al teletrabajo entre 2010 y 2022.



Gráfica 2: Comparación de términos relacionados al teletrabajo entre 2018 y 2022.

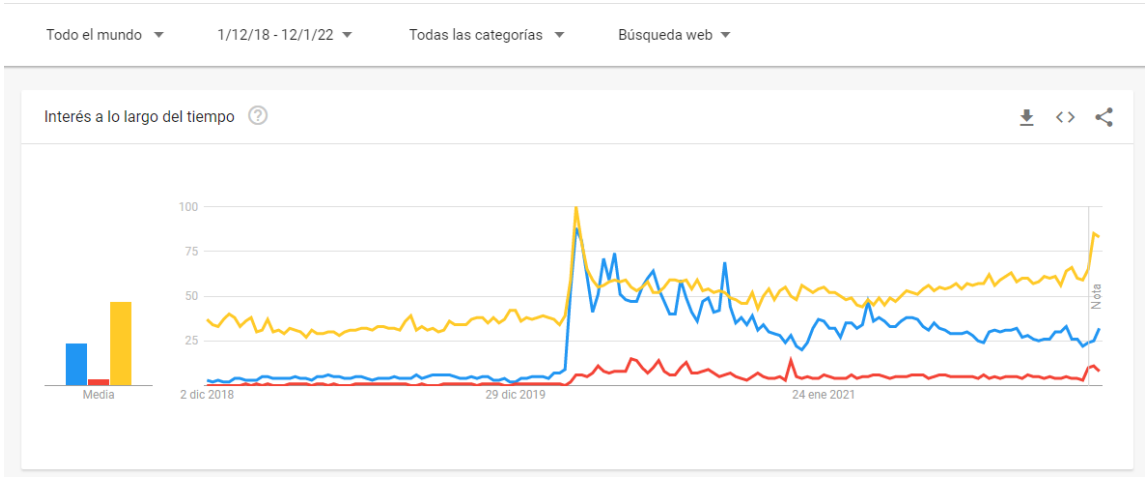


Tabla 1: Intereses de búsqueda de términos por región del mundo.