CIENCIA DE DATOS APLICADA A LA PLATAFORMA DE BENEFICIOS SOCIALES (PBS)

- INFORME FINAL -

02 de Noviembre de 2021

PROVINCIA DE CÓRDOBA

CONSEJO FEDERAL DE INVERSIONES

Secretaría de Fortalecimiento Institucional - Ministerio de Coordinación – Gobierno de la Provincia de Córdoba

Equipo técnico:

Manuel Infante

Contenido

I. Introducción

La aplicación de herramientas de Ciencia de Datos resulta vital para la promoción de políticas públicas de mayor eficacia y justicia social. En ese sendero, el Gobierno de la Provincia de Córdoba ha avanzado en la construcción de una Plataforma de Beneficios Sociales (PBS) que sintetiza información relevante y de acceso al programas sociales sobre un universo de personas localizadas en el territorio provincial.

Sobre esta base, se planificó el desarrollo del presente proyecto que tuvo por objetivo apoyar el proceso de trabajo computacional para fortalecer la idenitificación unívoca de personas y grupos familiares dispuestos en la PBS mediante la implementación de procesos que buscaban revisar y controlar la consistencia de los datos disponibles, el contraste de sus distintos orígenes y promover mecanismos que permitan generar estadísticas resúmenes útiles para la gestión provincial de programas sociales y asistenciales.

En ese sentido, se desarrollaron acciones tendientes a implementar procesos de revisión y control de la información disponible, el estudio de oportunidades de interoperabilidad de aplicativos y fuentes de información y una evaluación de cómo podrían mejorarse los procesos de aprovechamiento de datos para la toma de decisiones.

Este trabajo, desarrollado bajo un formato de asesoría e intercambio con los cuadros técnicos de la Secretaría de Fortalecimiento Institucional, es documentado en el presente informe final.

Las tareas desarrolladas buscaron, en forma particular, identificar oportunidades de mejora en la identificación únivoca de personas en la PBS y la delimitación más acertada de sus grupos familiares. Adicionalmente, realizar ejercicios de simulación y tratamiento de datos que permitieran testear cuáles combinaciones de atributos (variables identificadoras de ciudadanos) constituían un más acertado conjunto de variables de identificación particular en cada caso.

Por otra parte, se realizó un esfuerzo por delimitar y recomendar cuáles procedimientos de data science y machine learning podrían permitir realizar ejercicios

de simulación que permitieran evaluar y conformar grados de confiabilidad ante diferentes configuraciones de grupos familiares posibles, de acuerdo a un importante volumen y variedad de información disponible.

Finalmente, el trabajo pretendió realizar recomendaciones en materia de articulación a corto y mediano plazo entre la PBS, el procesamiento de dicha información mediante software especializados y las visualizaciones interactivas e interfaces útiles para la toma de decisiones en programas como Microstrategy (actualmente el Gobierno Provincial cuenta con licencia para su uso) u otros programas ligados a la aplicación de este tipo de herramientas.

De esta manera, el presente informe consta de diez apartados fundamentales, incluyendo esta breve introducción, destinados al abordaje de las diez tareas comprometidas en el Plan de Trabajo del Proyecto.

En primer lugar, en el apartado II se desarrollan lo diversos puntos concernientes al cumplimiento de lo comprometido en la Tarea 1, referida al diagnóstico de las variables identificadoras de personas utilizadas en PBS y posibles oportunidades de mejora en la identificación unívoca de las mismas.

Luego, en el apartado III se realiza un diagnóstico relativo a la actual delimitación de los Grupos Familiares en PBS (Tarea 2). Mientras tanto, el apartado IV se encuentra destinado a analizar las posibles oportunidades de mejora en la delimitación de Grupos Familiares mediante la compatibilización de diversas fuentes de información disponibles y a través del uso de herramientas de machine learning y ciencia de tratamiento de datos (Tarea 3).

Por su parte, el apartado V se corresponde con el análisis empírico de las distintas conformaciones de identificadores unívocos de personas y la extracción de conclusiones en este sentido (Tarea 4 y 5).

En el apartado VI se definen criterios para la aplicación de procedimientos de ciencia de datos para la evaluación de los grupos familiares en la PBS (Tarea 6), incluyendo nuevas herramientas desarrolladas para tal efecto a lo largo del plazo de ejecución del proyecto. En la misma línea, se esbozan recomendaciones para lograr una mejor

aproximación de los grupos familiares y la aplicación de herramientas de la ciencia de datos para una adecuada implementación y mejoramiento constante.

Asimismo, el apartado VII da cumplimiento al compromiso de evaluar el grado de confiabilidad de los grupos familiares bajo diferentes alternativas de conformación propuestas en el acápite anterior (Tarea 7), proponiendo en el apartado VIII sugerencias de adaptaciones adicionales posibles que podría aplicar el Gobierno Provincial a los fines de continuar indagando en esta temática.

En este sentido, se analizan los criterios que pueden ser tenidos en cuenta para obtener grupos familiares que brinden mayor confiabilidad respecto a la configuración actual (grupo único), de manera que pueda lograrse lo comprometido en la Tarea 8 del plan de trabajo.

Finalmente, el apartado IX se destina a presentar las conclusiones del trabajo diagnóstico realizado, evaluando diferentes estrategias de articulación informativa entre las fuentes de datos disponibles en la PBS, su procesamiento estadístico y la automatización de reportes con software de interfaces interactivas para el análisis de los resultados de indicadores descriptivos socio-demográficos.

En función de esto, en este último apartado se esbozan recomendaciones en materia de articulación y uso de la información disponible en la PBS, de acuerdo al diagnóstico expuesto en apartados previos (Tareas 9 y 10).

Además, se insta al lector a seguir algunos ejemplos de problemas y soluciones posibles en el siguiente repositorio.

A modo de cierre, los dos últimos títulos se encuentran destinados a la incorporación de las principales Referencias Bibliográficas consultadas para la producción de este material y la ejecución del proyecto, así como también a la exposición del Anexo en el que se incluyen las tablas correspondientes a los resultados de diagnósticos exploratorios trazados.

II. Diagnóstico relativo a variables identificadoras de personas utilizadas en PBS y posibles oportunidades de mejora en la identificación unívoca de personas

En esta primera tarea, se efectuará y documentará un diagnóstico de cuáles son las oportunidades de mejora en los procesos de identificación unívoca de personas en la PBS tomando en consideración los identificadores utilizados hasta el momento.

En base a esta delimitación, se presentarán propuestas concretas para abordar y evaluar escenarios alternativos de identificación de personas, de acuerdo a los insumos disponibles para tal efecto.

II.1. Variables Identificadoras de personas utilizadas en el Gobierno de la Provincia de Córdoba

En primera instancia, es necesario hacer referencia a cuales son las variables que se encuentran dentro del identificador o clave única de identificación de una persona, a partir de ahora denominado “pkid” (por sus términos en inglés “primary key”), con el que se trabaja en la bases de datos del Gobierno de la Provincia de Córdoba con el propósito de identificar de forma unívoca a una persona o individualizar al ciudadano. En la actualidad el pkid completo de una persona está conformado por las siguientes variables:

- id_sexo, referida al género de la persona (01, hombre y 02, mujer)
- pai_cod_pais, en relación al país del tipo de documento
- nro_documento, relativo al número de documento del ciudadano
- id_numero, campo identificatorio que va de 0 a 9 que permite complementar la identificación unívoca de una persona

II.2. Análisis de identificadores y tareas de depuración en el marco de la PBS

La plataforma de beneficios sociales se consolidará a partir de un proceso que implica la integración de distintas bases provenientes de diversos orígenes o fuentes de información, trabajo que comienza con el objetivo de almacenar la totalidad de los datos en una sola base de datos unificada con el propósito de potenciar la explotación de los mismos. Una vez consolidada y estandarizada, entre otras actividades, es posible unificar procesos de postulación, consolidar requisitos, desarrollar soluciones tecnológicas que permitan alcanzar la interoperabilidad de datos provenientes de distintas fuentes, unificar procesos de verificación de condiciones de admisibilidad y consolidar padrones de titulares de beneficios.

El punto de inicio del proceso es la recolección de las siguientes bases de datos:

- Padrón electoral
- Ministerio de trabajo
- Tarjeta social
- Tarjeta alimentar
- Pasivos
- Activos
- Gobierno
- Monotributo
- Vulnerabilidad Social

Previamente a la unión de las mismas, se lleva a cabo un proceso de depuración individual de cada base en términos de duplicados de acuerdo a las variables identificadoras que cada una de ellas contiene.

Luego se comienza a realizar la intersección de padrones o bases de datos, de forma secuencial. Al iniciar el mismo, quedan en evidencia dos problemas. En primer lugar, existen discrepancias en la disponibilidad de variables identificadoras en cada base de datos. Más precisamente, no todas cuentan con las 4 variables claves, muchas se encuentran sin información para una o más de estas. En segundo lugar, se observan situaciones en las cuales hay personas que su pkid difiere en alguno/s de sus campos,

según la fuente de información que se trate. A modo de ejemplo, la variable pai_cod_pais es aquella en la cual se encuentran mayores casos problemáticos debido a que alguno de los dos orígenes de información no presenta dato para este campo (es decir, se observa un valor “missing”).

En consecuencia, se generan registros duplicados a la hora de ir conformando la base unificada.

En cada etapa de inclusión de nuevas bases de información, es necesario llevar a cabo un minucioso proceso de análisis de la cantidad de registros y la calidad de los mismos, para, de ser necesario, realizar una depuración de la información resultante en la unión de los diferentes padrones, con el propósito de disminuir la duplicación de personas, que se es factible se genere y, minimizar errores.

Una vez depurada de la mejor forma posible la base que incluye todos los padrones anteriormente mencionados, se incorporan 3 bases que añaden información sobre las personas que ya han sido incluidas respecto a valuaciones, embarcaciones, provenientes de la Dirección General de Renta de la Provincia de Córdoba y datos que aporta SINTYS. En detalle, se incorporan los siguientes grupos de variables:

- Patrimoniales:
    1. Cantidad, antigüedad y valor monetario de automóviles
    2. Cantidad y valor monetario de Inmuebles
    3. Cantidad de embarcaciones
- Ingreso laboral o previsional:
    1. Jubilaciones y pensiones
    2. Asalariados registrados
    3. Autónomos
- Programas Sociales:
    4. Pensiones no contributivas
    5. AUH
    6. Otros programas sociales Nacionales, Provinciales y Municipales

- De caracterización:
    1. Discapacidad

A continuación, se incorporan las siguientes bases secuencialmente:

- Salas Cunas
- Paicor
- Gestión Educativa
- Mas Leche Mas Proteínas
- Tarifa Solidaria
- Boletos de transporte: Boleto Educativo Gratuito (BEG), Boleto Adulto Mayor (BAM), Boleto Obrero Social (BOS) y Boleto Social Cordobés (BSC). Dentro de estas bases también se incluyen las siguientes variables:
    1. Nivel Educativo estudiantes beneficiarios de BEG
    2. Nivel Educativo docentes beneficiarios de BEG
    3. Nivel Educativo Personal de Apoyo beneficiarios de BEG

Tabla Nº 1: Principales resultados en la incorporación de bases de Transporte

Programa	Nº observaciones	Variables identificadoras para reporte de duplicados	Cantidad de duplicados	Missing en variables identificatorias	Nº observaciones que cruzan con la base consolidada
BEG	227573	Id_sexo nro_doc pai_cod_pais id_numero	Si. 2	Si, id_sexo, id_numero, nro_doc, 4,678 registros y pai_cod_pais 4,683	210568
BAM	98471	Id_sexo nro_doc pai_cod_pais id_numero	Si. 6	Si, id_sexo id_numero pai_cod_pais y nro_doc, 78 registros	94164
BOS	16960	Id_sexo nro_doc pai_cod_pais id_numero	Si. 3		15211
BSC	11002	Id_sexo nro_doc pai_cod_pais id_numero	Si. 1		10636

El siguiente paso es incorporar las bases referidas a personas en situación de vulnerabilidad1, detectadas por 3 fuentes de información diferentes:

- Relevamientos de asistentes sociales en Córdoba Capital (1)
- Base de Municipios y Comunas (2)
- Base decilinferior Julio (3)

Luego, se agrega base referida a Monotributo Social con su respectiva variable referida a letra de monotributo.

Posteriormente se incluye las bases de los programas pertenecientes al Ministerio de la Mujer del Gobierno de Córdoba y se realiza el análisis de duplicados pertinente. Las mismas son:

- Nueva Vida
- Protección de la embarazada y su bebé
- Nuevo rumbo – Alquileres
- Nuevo rumbo - Subsistencia
- Cobertura de Salud mujeres

Finalmente se adicionan a la base consolidada las bases:

- Vida Digna.

A modo de síntesis, la siguiente tabla expone el listado completo de bases participantes del proceso de integración de padrones y la disponibilidad de las variables necesarias para el logro de la identificación univoca de las personas:

[1] Definición de vulnerabilidad:

Tabla Nº 2: Disponibilidad de variables identificatorias (pkid) según fuente de información

Nº	Nombre de la base	Id_sexo	Nro_documento	Pai_cod_pais	Id_numero
1	Padrón Electoral	SI	SI	SI	SI
2	Ministerio de trabajo	SI	SI	SI	SI
3	Tarjeta Alimentar	SI	SI	SI	SI
4	Tarjeta Social	SI	SI	SI	SI
5	Pasivos	SI	SI	SI	NO
6	Activos	SI	SI	SI	NO
7	Gobierno	SI	SI	SI	NO
8	Monotributo	NO	NO	NO	NO
9	Vulnerabilidad Social	NO	NO	NO	NO
10	Salas Cunas	SI	SI	SI	SI
11	Paicor	SI	SI	SI	SI
12	Gestión Educativa	SI	SI	SI	SI
13	Mas Leche Mas Proteínas	SI	SI	SI	SI
14	Tarifa Solidaria	SI	SI	SI	SI
15	BEG	SI	SI	SI	SI
16	BAM	SI	SI	SI	SI
17	BOS	SI	SI	SI	SI
18	BSC	SI	SI	SI	SI
19	Vulnerabilidad_(1)	SI	SI	SI	SI
20	Vulnerabilidad_(2)	NO	SI	NO	NO
21	Vulnerabilidad_(3)	NO	NO	NO	NO
22	Vida Digna	NO	SI	NO	NO
23	Ministerio de la Mujer	NO	SI	NO	NO

Cabe destacar, como se observa en las tablas expuestas, que es factible que para algunas de las bases existan registros que no traigan el valor o dato asociado a cierta variable o campo, es decir, que presenten “missing” en alguna de las 4 variables del pkid.

Una vez alcanzada la consolidación de la base, considerando la inclusión del total de los orígenes de información disponibles hasta el momento, para proseguir con el desarrollo del diagnóstico y el reconocimiento de oportunidades de mejora de la identificación de personas, en primer lugar, es necesario llevar a cabo una serie de actividades que permiten una mejor y correcta adecuación de esta base de datos utilizada como insumo principal de la PBS.

II.3. Adecuación de la base de datos

II.3.1 Proceso de recuperación de datos

Como se explicó anteriormente, la presencia de datos faltantes en alguna de las variables del pkid genera ciertos problemas en la identificación de las personas, principalmente, la generación de duplicados y la imposibilidad de asignarle ciertas cualidades personales a los registros, tal como su condición laboral, posesión patrimonial, beneficios de programas sociales y características generales de la persona.

Siguiendo el objetivo de mejorar la identificación de las personas se procede de la siguiente forma:

En primer lugar, se analiza la cantidad de datos “missing” de la base consolidada en cuanto variables “cuil”, “nro_documento” y “id_sexo”. La siguiente tabla muestra la cantidad total para cada caso:

Tabla Nº 3: Datos missing según variable de interes analizada respecto al total de la base

Variable	Cantidad de datos “missing”	Participación porcentual (%) en el total de la base
CUIL	243.232	5,74%
Nro_documento	46..776	1,10%
Id_sexo	76320	1,80%
Total de registros	4.239.693	

Posteriormente, se cruza la base consolidada, por medio de SQL, con bases provistas por la Dirección de Ciencia de Datos poniendo esencial atención en las variables

cuil_rc, id_sexo_rc, nro_documento_rc, pai_cod_pais_rc, id_numero_rc, nombre_rc, apellido_rc y fec_nacimiento_rc, las cuales nos permiten mejorar el análisis de identificación. De forma siguiente, se eliminan el conjunto de variables irrelevantes para nuestro análisis, provenientes de las bases enviadas por la Dirección mencionada al inicio de este párrafo.

Luego, se realiza un proceso de recuperación de datos para de cada una de las variables mencionadas como relevantes en el párrafo anterior y se ve que, la cantidad de datos perdidos presentan una notable mejoría. A continuación, en la tabla N°4 se puede observar, para las tres variables presentadas en la Tabla 3, que sus totales descienden.

Tabla Nº 4: Presencia de datos faltantes

Variable	Cantidad de datos “missing”	Participación porcentual (%) en el total de la base
CUIL	147.866	3,49%
Nro_documento	29	0,00%
Id_sexo	944	0,00%
Total de registros	4.239.693	

Aquí se puede apreciar una baja significativa de datos “missing” en cada variable analizada. En particular, vemos que nro_documento contiene solo 29 missing, mientras que id_sexo tiene un poco menos de 1000 datos perdidos. Finalmente, al observar la variable cuil, si bien se da una baja sustancial, quedan aún unos 150.000 casos sin dato recuperado.

Dado el hecho de que se cuenta con campos más completos en las variables id_sexo, nro_documento y cuil, cabe remarcar que en esta instancia aparecerán más duplicados en término de estas variables, ya que se tiene la posibilidad de realizar un análisis de duplicados a partir de variables con menor cantidad de “missing”. Se contempla la posibilidad de realizar este tipo de tareas en etapas subsiguientes del proceso de depuración y mejora de la base de datos.

Debido al hecho que se tenía a disposición información sobre fecha de nacimiento de las bases provenientes de la Dirección de Ciencias de Datos, se opta por completar

con esta información aquellos registros con missing en esta misma variable en la base DEA, y a partir de la variable fecha de nacimiento ya completa, se calcula la variable "edad_DEA" para aquellos registros que presentaban información. Los resultados se sintetizan en la siguiente tabla:

Tabla Nº 5: Variable edad a completar

Análisis de Edad_DEA a completar	Cantidad
Cantidad missing en Edad_DEA	68.479
Cantidad datos en fec_nac_rc y RC_fec_nac	4.127.919
Cantidad missing en Edad_DEA pero completo en fec_nac_rc y RC_fec_nac	45.952

Observando estos resultados, se espera que se puedan completar aproximadamente 45.000 mil registros con edad, quedando sin completar más o menos unos 23.000. Efectivamente, en concordancia con lo esperado, al calcular la edad_DEA en base a fec_nac_rc, nos quedan unos 22.829 sin datos en esa variable.

II.4. Primera aproximazión al problema de duplicados

II.4.1 Diagnóstico preliminar

Se comienza por observar en la siguiente tabla cómo se comportan los duplicados bajo el criterio de identificación por DNI por un lado, y por otro, DNI y CUIL, además de evaluar su incidencia según si las observaciones están o no en el padrón.

Tabla Nº 6: Duplicados por DNI, CUIL dentro y fuera del padrón

Selecciona un área para agregar un comentario

Reporte de duplicados	Universo	Duplicado por Criterio			
		DNI	DNI y CUIL	% duplicados DNI	% duplicados DNI - CUIL
En padrón electoral	2,885,662	12,718	226	0.44%	0.008%
Fuera de padrón electoral	1,307,255	29,726	10,352	2.27%	0.792%
Casos con DNI no missing	4,192,917	102,606	23,376	2.45%	0.558%

Cálculos sin considerar DNI missing en base a BUC marzo_junio 2020

Es claro que cae prácticamente a cero el porcentaje de duplicados en padrón electoral cuando se utilizan CUIL y DNI, a la vez es importante resaltar que hay una mayor incidencia de duplicados fuera del padrón incluso cuando se trabaja sin incluir los casos con dni missing.

II.4.2 Identificación de duplicados por fuente de información

Para la identificación por medio de fuentes de información adjuntamos a continuación una tabla con los duplicados observados según la fuente de datos.

Tabla Nº 7: Duplicados según fuente de información

CANTIDAD DE DUPLICADOS SEGÚN FUENTE DE INFORMACIÓN				
Duplicados DNI por fuente de datos	**Padrón Electoral**	**Fuera de Padrón Electoral**	**Total**	**% de cada fuente de datos**
Vida Digna	829	3.961	4.790	5,4%
Ministerio de la Mujer	52	82	134	3,7%
Tarjeta Alimentar	1.383	1.910	3.293	2,8%
Transporte	4.370	4.195	8.565	2,4%
Vulnerable MyC	869	554	1.423	2,4%
Tarjeta Social	908	534	1.442	2,2%
Caja de jubilaciones	4.159	1.992	6.151	2,1%
Monotrib Sociales	114	169	283	2,0%
Tarifa Solidaria	2.844	1.055	3.899	1,9%
Monotributo	3.142	1.941	5.083	1,6%
IFE	6.080	3.299	9.379	1,5%
Policía Fiscal	14.227	48	14.275	1,3%
Gestion Educativa	1.235	8.910	10.145	1,1%
Mas leche mas proteína	1.184	274	1.458	0,9%
Vulnerable decilinferior	881	34	915	0,9%
Paicor	31	2.191	2.222	0,8%
Salas Cuna	1	0	1	0,0%
Totales	**42.309**	**31.149**	**73.458**	

Cabe destacar que es importante contextualizar cada fuente de información con la cantidad de observaciones que ésta produce, ya que de esta forma se observa que aquellas fuentes que presentan la mayor cantidad de son las mismas que brindan más información, por ende, no son las más ineficientes en términos de duplicados por cantidad de observaciones.

II.4.3. Propuesta de identificadores

Además de la propuesta básica de utilizar a variables identificadoras al DNI y CUIL es factible considerar adicionar la variable fecha de nacimiento. Todas estas variables tienen como principal ventaja que no cambian en el tiempo, si lo hacen son casos excepcionales.

Los problemas subyacentes a estas variables consisten en que aún se requiere recuperar datos faltantes y una nueva limpieza para atenuar casos críticos, además de que, para determinadas tareas, como cruces informativos se deberá seguir trabajando con el PKID (aquí reside la importancia de la mantención de su calidad).

También debe considerarse utilizar a las variables del PKID pero despreciando o permitiendo diferencias sólo en nacionalidad.

II.4.4. Nuevas oportunidades de mejora: Identificación de variables clave mediante método iterativo

Para la presente sección se tiene como objetivo la búsqueda de conjuntos de variables que se presenten como buenos identificadores y técnicas más eficientes por medio de herramientas de ciencias de datos para complementar aquellas planteadas en los apartados anteriores. Entre ellas se propone:

- Concebir bases como objetos relacionales con atributos asociadas e interdependencia entre ellas. Con respecto a lo primero, atributos serían periodo de corte de las observaciones, fuente, formato, ubicación, metodología, entre otros. Por el otro lado, cuando me refiero a interdependencia hago alusión a esquemas de bases de datos.[2]
- Armar correspondencias con expresiones regulares que ayuden en la depuración e intersección de los datos [3]
- Establecer un orden de prioridades/preferencia entre fuentes para hacer el cruce de tablas.
- Identificación de variables clave mediante una función de perdida calculada en base a un método iterativo.[4]

---

[2] Esquema de Basse de Datos: El esquema de una base de datos (en inglés, database schema) describe la estructura de una base de datos, en un lenguaje formal soportado por un Sistema de Gestión de Bases de Datos (DBMS). En una base de datos relacional, el esquema define sus tablas, sus campos en cada tabla y las relaciones entre cada campo y cada tabla.

[3] Ver Ejemplo 1 en el repositorio.

[4] Los pasos a seguir resumidamente serían:

1) Colocar la condición de exclusión de las variables id_numero, pai_cod_pais, id_sexo.

2) Luego, mediante el uso de “itertools” (paquete de pandas Python) se realiza la iteración para determinar la combinación de variables óptimas

III. Diagnóstico relativo a la actual delimitación de Grupos Familiares en PBS

A los fines de analizar las estructuras de delimitación de grupos familiares que se disponen en PBS, sus orígenes informativos y grado de actualización general y particular; se presenta a continuación un diagnóstico relativo a la evolución observada en la definición de Grupo Familiar hasta llegar a la actual concepción de Grupo Único o Conviviente.

De esta manera se pretende exponer las particularidades de cada definición y analizar las fortalezas y debilidades existentes. Lograr la delimitación más acertada posible de los grupos familiares es de suma importancia debido, entre otros aspectos, a que las condiciones de elegibilidad vigentes en los principales programas sociales de la provincia involucran una evaluación de la situación socio-económica del grupo familiar del postulante.

III.1. Grupo familiar

En una primera instancia, en el año 2004, se comienza a trabajar con el concepto de "grupo familiar" a solicitud de poder establecer un grupo familiar para cada persona.

Se comienza con la definición de grupo familiar genérico, que tiene asociada la característica de tener un vínculo consanguíneo patriarcal. En cada grupo familiar, al padre de familia o madre de familia (de no existir un padre de familia) se le otorga la figura de "Jefe de Familia".

Dada la particularidad que presenta cada programa social o asistencial desarrollado en el ámbito del Gobierno de la Provincia de Córdoba, queda en evidencia que existen

3) La salida de sistema presenta para cada conjunto de variables la cantidad de duplicados e identificados

4) En base a ello se calcula la efectividad de dicho conjunto de variables siendo esta la que es trascendental en la elección de las variables óptimas.

casos en los cuales una aplicación puede precisar disponer de un grupo familiar diferente al convencional (donde se entiende que el jefe de familia es el padre o madre), entonces se observa que se comienzan a armar otros grupos familiares propios, de manera independiente al genérico.

A modo de ejemplo, programas como PAICor, Salas Cuna y Tarjeta Social han tenido conformaciones de grupos familiares específicas, que han podido ser unificadas progresivamente a través de la implementación del Formulario Único de Postulantes (FUP) como mecanismo de acceso a los programas ejecutados por el Gobierno Provincial.

En este sentido, el esquema de "grupo familiar" aún sigue vigente y está en funcionamiento para aquellas aplicaciones que aún no han realizado el pasaje o migración a Grupo Único5 (concepto utilizado en la actualidad).

En este caso, para su diferenciación, la finalización del Id del grupo familiar indica si se trata de un grupo genérico (00) o de alguna aplicación (por ejemplo en el caso de PAICor es el 13).

Sin embargo, luego de muchos años de coexistir múltiples grupos familiares con distintas figuras y vínculos, según las diferentes aplicaciones, es que se comienza a advertir ciertas inconsistencias y actitudes oportunistas en la declaración de grupo realizada. Los grupos familiares se declaran y modifican acorde a los requerimientos de cada programa social o asistencial con el objetivo de ser parte del conjunto de personas que cumple con los requisitos para ser beneficiario del mismo.

Siendo así, en el año 2017 se empieza a pensar en un nuevo esquema de "Grupo Único Conviviente" como proyecto, para superar las dificultades encontradas en el esquema anterior.

[5] Desarrollado en el próximo apartado

III.2. Grupo conviviente o grupo único

Entonces, el concepto de Grupo Familiar migró o evolucionó a Grupo único. Este último es sinónimo de Grupo Conviviente. Se entiende como convivientes a un grupo de personas que viven bajo un mismo techo.

En comparación con grupo familiar, grupo único ha dejado de lado la cuestión sanguínea, ya no se tiene en consideración esta relación al momento de la conformación del mismo. No existe esta vinculación entre las personas.

En virtud de ciertos requerimientos o condiciones de admisión para programas sociales o asistenciales puede solicitarse el levantamiento de la relación sanguínea entre conformantes de un mismo grupo único pero con el objetivo de corroborar alguna relación de interés o relevancia para cierto análisis. Sin embargo, no es posible modificarla o asociarla a grupo único. Actualmente, solo Registro Civil tiene la potestad para cargar esta clase de datos.

En este sentido, una persona puede estar en un solo grupo único a la vez y de esta manera, se evita el armado de un grupo a conveniencia para recibir algún beneficio otorgado por el estado.

Entonces con la nueva delimitación, surge la aplicación de Grupo Único y la definición de IDVIN, el cual determina una identificación de domicilio único. El IDVIN implica un valor único que identifica atributos geográficos como ser departamento, localidad, barrio, calle, altura única. Es decir, hay un conjunto de personas nucleadas en un domicilio único, IDVIN.

Al pertenecer a un grupo, de forma obligatoria y automática, se debe tener un domicilio. No obstante, puede suceder que un IDVIN le corresponda a dos grupos únicos distintos (con personas distintas). Es de público conocimiento que en ocasiones, dos “familias” o “grupos convivientes” comparten la vivienda y se hospedan bajo el mismo techo.

Otra característica es que dentro de la aplicación de grupo único existen ciertos programas que tienen la potestad de modificar el domicilio de grupo único, pero no sucede en todos los casos. Algunos solo pueden leer y no editar.

Es importante destacar que existen distintas clases de domicilios para una persona:

- Domicilio legal: el que figura en el DNI, el único organismo que puede modificarlo es RENAPER
- Domicilio real: aquel declarado por la persona, lo puede cambiar cualquier ejecutor de programa que tenga esa potestad. Este domicilio tiene la particularidad que se va actualizando al declarar domicilios diferentes.
- Domicilio de grupo conviviente/único: declarado para grupo conviviente.

En el caso de que un ciudadano declare un domicilio para la aplicación grupo único que difiera del domicilio real, automáticamente se actualizará el domicilio real (para todas las personas que forman parte de ese grupo conviviente) a aquel declarado en grupo único. No obstante, no sucede lo mismo a la inversa.

Una nueva particularidad de la aplicación grupo único es que el ciudadano tiene la posibilidad de modificar su grupo declarado. Anteriormente, estos cambios estaban a cargo de trabajadores de gobierno, no se encontraba abierto al ciudadano. De querer modificar (ya sea incorporando a una nueva persona o eliminando a algún miembro del grupo), la aplicación advertirá que uno se encuentra próximo a modificar o separar al grupo único (de corresponder) y una vez aceptado el cambio no es posible deshacerlo.

En relación a este último punto, actualmente se están analizando ajustes frente al desempeño que tiene la aplicación en lo relativo al uso que le da el ciudadano. Sucede, a modo de ejemplo, que hay personas que están sin grupo único, ya que puede haberse eliminado (o haber sido eliminado) de algún grupo previo en el que se encontraba. Un objetivo a corto plazo es lograr armar grupos unipersonales, con asignación automática, para este subconjunto de personas, ya que es una realidad observada la existencia de éstos.

Por último, existe la posibilidad de ver el histórico de los grupos para ver los cambios en el tiempo y los comportamientos o actitudes de la población. Esto último tiene relevancia para los tomadores de decisiones en relación al otorgamiento de programas o beneficios sociales, créditos bancarios, entre otros. Todas las decisiones de gestión se están tomando alrededor del concepto de grupo único, por eso la importancia de su correcta determinación.

Sin lugar a dudas resta mucho trabajo por realizar y es necesario aprovechar toda herramienta de ciencia de datos disponible para continuar avanzando en la calidad de esta dimensión analizada.

En definitiva, parte del trabajo a realizar es lograr unificar e integrar las diferentes definiciones de “grupo” que siguen conviviendo, bajo algún criterio de ordenamiento lógico y poder plasmar todo en un solo concepto mejorado y enriquecedor.

IV. Análisis de posibles oportunidades de mejora en la delimitación de Grupos Familiares

En función de lo realizado en la tarea 3, se evaluaron los insumos de información disponibles que podrían aportar referencias útiles para la evaluación de una redefinición de grupos familiares.

En particular, el análisis realizado buscó la compatibilización de las diversas fuentes de información disponibles que pudieran favorecer la delimitación correcta de grupos familiares con el uso de herramientas vigentes de machine learning y ciencia de datos.

De esta manera, se pudieron trazar diagnósticos acerca de cómo podría abordarse este aspecto, a través de la utilización de ejercicios de simulación de distintas construcciones de grupos familiares.

A juzgar por el estado actual de la definición de grupos convivientes, se considera que el principal problema es la identificación univoca de cada persona. Luego de salvado este problema con las técnicas mencionadas, se considera pertinente avanzar con la siguiente lógica:

1. Evaluar diferencias si luego de identificar correctamente a cada persona los grupos convivientes siguen siendo representativos con métricas de distancia.
2. Explorar la situación de domicilios desde una perspectiva de Georeferenciación *hacia adelante*[6]
3. Estructurar un esquema de monitoreo para la evolución futura de los Grupos Convivientes con metodologías de Clusters y Clasificación.

A lo largo del plazo de ejecución del Proyecto se realizaron intercambios con cuadros técnicos del Gobierno Provincial en los que se pudo trabajar sobre los puntos señalados, con resultados y perspectivas alentadoras.

[6] Ver Ejemplo 2 en el repositorio.

En particular, se creó una lógica de consultas al proveedor de datos LocationIQ que permite transformar datos estandarizados de la forma habitual en las bases administrativas a Objectos de JavaScipt que poseen clasificadores extremadamente útiles. Para clarificar estos términos, plasmaremos un ejemplo:

Pongamos por caso una dirección aleatoria de una base de datos de domicilio enviada por Sintys, sin ninguna otra información adicional. En particular, consideramos un caso sin ningún tipo de procesamiento, con excepción de una concatenación de campos:

Consulta efectuada: "Entre Rios,2069,2659,Monte Maiz,Union,Cordoba, Argentina"

Respuesta:


```
[
 {
  "place_id": "72792494",
  "licence": "https://locationiq.com/attribution",
  "osm_type": "way",
  "osm_id": "387350887",
  "boundingbox": [
   "-33.206494763265",
   "-33.206394763265",
   "-62.597609361224",
   "-62.597509361224"
  ],
  "lat": "-33.2064447632653",
  "lon": "-62.59755936122449",
  "display_name": "2069, Entre Ríos, Monte Maíz, Municipio de Monte Maíz, Pedanía Ascasubi, Departamento Unión, Córdoba, X2550, Argentina",
  "class": "place",
  "type": "house",
  "importance": 0.42099999999999993
 } ]
```

En la respuesta apreciamos en primer lugar las coordenadas (lat, lon), que se encuentran dentro de un área de búsqueda (componente bounding box) configurable.

En segundo lugar vemos la clasificación dentro del Modelado de Datos, está categorizada como “camino” (way) con una clasificación de “lugar-hogar” (“place-house”)[7]. En tercer lugar, tenemos un parámetro con la confianza del resultado (“importance”) que nos permite decidir en caso de tener más de un resultado.

Con este ejemplo queremos mostrar que es posible enriquecer las fuentes de información que se tienen hasta el momento, a un coste nulo. Al respecto, se observa que al contar con información y fuentes de datos alternativos en lo que respecta a la conformación de variables de domicilio, que además en general se encuentra estandarizada, se tiene una ventaja interesante para avanzar en el proceso de georreferenciación de individuos y grupos familiares de la PBS.

Esto abre la puerta para poder realizar futuros análisis de distribución geo-espacial de características socio-demográficas y de percepción de programas sociales, lo que permite aplicar herramientas de clusters, clasificación y análisis geográfico, con la posibilidad de dar seguimiento a esta información en el tiempo, combinando datos espaciales y en serie temporal como insumo para la gestión.

En esta línea, las recomendaciones vertidas a equipos técnicos de gobierno permitirán la evaluación del uso de este tipo de instrumentos a una mayor escala, especialmente para la delimitación de la ubicación geográfica y la conformación del grupo familiar más adecuado para la población objetivo de la PBS, principalmente focalizada en poblaciones vulnerables y en situación de riesgo social.

[7] El Modelo cuenta con nodos, caminos y relaciones, donde cada uno tiene una clasificación. Para ampliar sobre este punto visitar el sitio del proyecto OpenStreetMap

V. Simulación y evaluación de duplicados en la PBS según distintos criterios.

El objetivo de la presente sección reside en el contraste empírico de distintas estructuras de variables identificadores, para ello se plantea en cada subsección una estructura y descripción simplificada.

V.1. Identificación de duplicados por disponibilidad de PKID.

Retomando las variables que conforman el PKID (id_sexo, pai_cod_pais, nro_documento, id_numero) realizamos un nuevo análisis de duplicados tomando a este conjunto de variables identificadoras unívocas de personas, notamos en la siguiente tabla (Tabla 8) que existe una mayor preponderancia de duplicados en bases de datos con PKID incompleto.

Tabla Nº 8: Duplicados según PKID

	pkid comp	pkid incomp
Cantidad duplicados	50,048	23,410
Cantidad de registros totales s/fuentes	3,490,717	1,306,615
Tasa de incidencia	1.4%	1.8%

V.2. Duplicados bajo Cuil, DNI y fecha de nacimiento.

Como contraste empírico de las variables propuestas en base a la última versión de la base en servidor de Gobierno y aplicando los criterios de identificación bajo el siguiente conjunto de variables: “cuil”, “nro_documento” y fecha de nacimiento, se pueden observar 4,116 millones de registros sin duplicados sin embargo aún persiste 111205 registros con problemas de duplicación, a continuación, se adjunta una tabla en concepto de resumen.

Tabla Nº 9: Duplicados según cantidad de copias

copias	observaciones	excedente
1	4,116,002	0
2	83,038	41,519
3	24,492	16,328
4	480	360
5	15	12
6	12	10
7	35	30
9	9	8
15	30	28
3094	3,094	3,093

V.3. Duplicados bajo base depurada.

Operando sobre la base resultante del inciso II.3, se han identificado la cantidad de registros con duplicados en términos de “nro_documento” y “cuil”. En la siguiente tabla se exhiben los resultados:

Tabla Nº 10: Duplicados en términos de ″nro_documento” y ″cuil”

Duplicates in terms of nro_documento cuil

copies	observations	surplus
1	4137195	0
2	79148	39574
3	22836	15224
4	484	363
5	5	4
26	26	25

Tenemos un total de 102.499 registros con duplicados, sobre estos se han identificado aquellos en términos de las variables “id_sexo”, “nro_documento” y “cuil” siendo un total de 75.287 registros con duplicados. A estos se les ha reemplazado la información en las filas “sobrantes” y se han eliminado los duplicados, dropeándose un total de 41.386 registros. Sobre los registros que quedaron se han identificado aquellos que tienen mismo “nro_documento” y “cuil”, pero diferente sexo, se trata de un total de 26.587 registros, los cuales finalmente se cruzaron con la base de Renaper, verificándose el sexo correcto de un total de 11.592 personas.

Entonces, del total de registros con duplicados (102.499) al eliminarse la primera tanda de duplicados e incluyendo la eliminación de registros duplicados con sexo

incorrecto, nos queda un total de 46 mil registros para sumar a la base master. Antes de sumarlos se eliminaron los 102.499 registros duplicados.

Una vez incorporado los 46 mil registros, nos queda una base de 4.183.313 registros, el resumen de las cantidades de missing por variables id se presentan en la siguiente tabla:

Tabla Nº 11: Faltantes por variable en base actualizada

Variable	Cant_missing
id_sexo	944
nro_docume	28
cuil	145,867
fec_nacimie	109,972
edad_DEA	21,037

Sobre esta base final, se presentan los resultados del reporte de duplicados en la próxima tabla:

Tabla Nº 12: Duplicados en términos de "nro_documento" y "cuil" en base actualizada

copies	observations	surplus
1	4182282	0
2	1006	503
11	11	10
14	14	13

En conclusión, pudo resolverse el problema de duplicados con criterios documentados que hacen posible volver a enfrentar el problema para nuevas fuentes de información.

A colación de esto, se elaboraron rutinas de trabajo que permitieran abordar esta cuestión de manera automática, utilizando los pasos descriptos arriba.

VI. Aplicar ciencia de datos para la determinación de grupos familiares en la PBS

En el presente apartado como objetivo principal se busca la determinación de los grupos familiares de la manera más efectiva, para ello se presenta el desarrollo de nuevas herramientas las cuales pueden ser reutilizadas en otros proyectos, posteriormente se aproximan las posibles soluciones, luego se implementó la solución más plausible en la determinación de los grupos familiares en la PBS.

Por otro lado, se incluyó de manera adicional el análisis y procesamiento de imágenes de partidas de nacimiento para la extracción de información relativa al nacido y sus progenitores.

VI.1. Nuevas herramientas

En ánimos de aprovechar la base obtenida, se buscó unificar las conexiones utilizando el entorno de desarrollo “Jupyter Lab”. La elección se debe a que esta herramienta brinda la posibilidad de utilizar Python para analizar datos de forma interactiva en instancias llamadas “Notebooks” y a un coste nulo.

De esta forma, en primer lugar, se pueden utilizar módulos externos de procedimientos para unificar todas las conexiones de datos que la Plataforma de Beneficios Sociales utiliza en la actualidad. Por otro lado, bajo el mismo entorno se pueden ejecutar comandos en Stata, SQL y R lo que nos permite utilizar rutinas que ya se vienen utilizando y aprovechar toda la potencialidad de la interacción entre ellas. Finalmente, las métricas calculadas pueden presentarse al consumidor de la información con gráficos interactivos de manera sencilla.

En otras palabras, bajo un mismo entorno de ejecución se tiene a disposición las distintas fuentes de información y las numerosas herramientas de análisis que se ofrecen en múltiples plataformas de desarrollo.

Con respecto a las conexiones, para tener a disposición las distintas fuentes de información, se identificaron las fuentes de datos habituales:

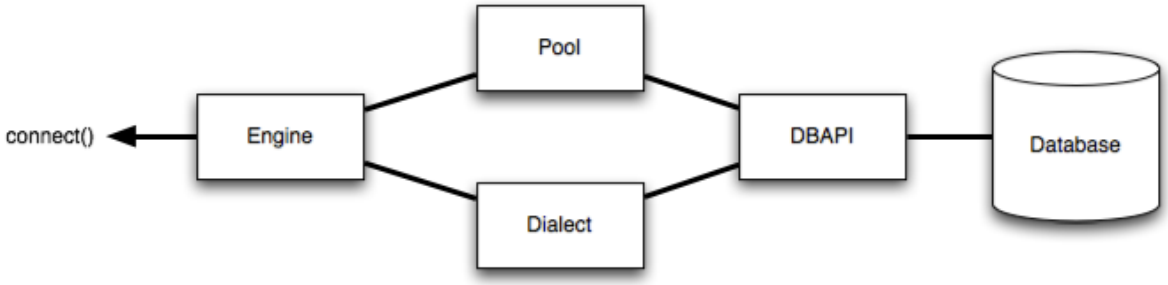
- Libros de Cálculo de Microsoft Excel y Tablas en text plano formato CSV (Comma Separated Values)

  La estrategia es estandarizar los cómputos habituales utilizando la framework de análisis de Pandas.

- Bases de Datos en un servicio de ORACLE.

  Aquí se utiliza el esquema de ORM (Object Relational Mapper) con el kit de herramientas de SQLAlchemy. Más precisamente, se crea un motor en SQLAlchemy que utiliza la DBAPi de cx_Oracle. O gráficamente:

Ilustración 1: Esquema de conexión a Base de Datos con SQLAlchemy



Luego de creado el motor se pueden ejecutar consultas utilizando la lógica de SQL y exportar los resultados a objetos de la clase Pandas DataFrame.

El objetivo de este desarrollo es lograr con pocas líneas de comandos la creación de procesos que integren todas las fuentes de datos y aprovechen la flexibilidad que un lenguaje multipropósito de alto nivel como Python brinda.

En cuanto al procesamiento, como se mencionó anteriormente, se buscó integrar dentro del esquema 3 alternativas:

Cómputos en R utilizando la interfaz rpy2.

- Cómputos en Stata con procesamiento por lotes utilizando la librería ipystata.

  A los fines de hacer más ameno el proceso de transición para el grupo de trabajo se configuró el protocolo ODBC en todos los equipos utilizados para poder acceder a la Base de Datos de Oracle desde la interfaz habitual de Stata.

- Finalmente, se presentó una alternativa para trabajar con grandes volúmenes de datos utilizando la librería de datos Vaex, atendiendo a la limitación de contar solo con un equipo para el procesamiento.

VI.2. Mejorar la aproximación

Retomando lo planteado en el inciso 4, considerando la obtención de una mejora consistente en materia de duplicados, es primordial poder detectar cuál es el domicilio correcto de cada individuo para la determinación de su grupo conviviente.

Para ello se utilizó un conjunto de datos enviados por SINTyS (Sistema de Identificación Nacional Tributario y Social).

De esta forma llegamos a que sintéticamente la base se puede concebir como:

$$Dataset = [personas\ domicilios]$$

Donde las informaciones relativa a las personas se concentran con las siguientes variables:

$$personas = [id\_persona, id\_sexo, nro\_documento, pai\_cod\_pais, id\_numero, apellido\_orig, nombre\_orig, cuil, fec\_nac, id\_grupo\_unico]$$

Mientras que la información correspondiente a los domicilios se encuentra en las siguientes variables:

$$domicilios = [id\_grupo\_unico, id\_domicilio, base\_origen, calle, nro, piso, depto, codigo\_postal, localidad, departamento, provincia]$$

Y actualmente la relación entre personas y domicilios es la siguiente:

$$personas\ 1 : n\ domicilios$$

Pero si el objetivo es detectar donde vive cada persona se prioriza la búsqueda de una relación:

$$personas\ 1 : 1\ domicilios$$

Para ello se plantearon dos alternativas:

1. Elegir en base orden de preferencia de acuerdo a la fuente de información que dio origen al domicilio. Un ejemplo puede ser la marca temporal de recolección del punto maestral.
2. Utilizar el domicilio que prevalece sobre la mayor cantidad de domicilios disponibles para esa persona.

Para la primera alternativa contamos con 7 fuentes de información:

Tabla Nª 13: Frecuencia de domicilios según fuente de información

Fuentes de información	Frecuencias	
AFIP	4140243	0.350865
ELECTORAL	2942210	0.249338
RENAPER	2646204	0.224253
OTRAS FUENTES - BNA	1295174	0.109760
ENARGAS	577423	0.048934
missing	161719	0.013705
OTRAS FUENTES - BC	24514	0.002077
ENRE	12614	0.001069

Sin embargo, no se cuenta con una marca temporal para establecer un orden de preferencia y a la vez tampoco es real asumir que todas las observaciones que provienen de una fuente fueron registradas al mismo tiempo. Es por ello que se decidió implementar la segunda solución.

Es decir, se busca el conjunto de D que incluye las direcciones

di para cada individuo i tal que:

$$D = \{ \quad d_i : max\ f(d) \text{ para todo } i \quad \}$$

$$i \in [1, n] \text{ (personas)}$$
$$d_{1xk} \text{es el vector de k direcciones posibles para cada individuo } i$$
$$f \text{ es la función de frecuencias}$$
$$f : R^n \rightarrow R^n$$

$$max(\ f(d)\ ) : R^n \rightarrow R$$

Notar que no hay nada que nos garantice que ese máximo sea único. Y, por otro lado, que la dimensión del vector d depende de la cantidad de domicilios registrados para cada persona es variable y no tiene por qué coincidir con el número de fuentes de información.

VI.3. Implementación.

Los pasos necesarios para la determinación de los domicilios correctos son los siguientes:

1. Cambio de formato y eliminación de malformaciones

Los datos se encontraban en un formato de texto plano delimitados por un pipe o "|" con errores de formato en algunas observaciones. Para corregir esto se utilizó un método de "try…catch" para escanear los errores y luego una conversión a formato hdf5 por bloques.

2. Geolocalización de direcciones.

Dado que la implementación del operador igualdad en secuencia de caracteres presenta muchos inconvenientes, se decidió por trabajar sobre pares de coordenadas aprovechando los algoritmos de aproximación que las APIs de localización utilizan.

Más precisamente, se inició un servidor EC2 gratuito en AWS (Amazon Web Services) desde el cual se hicieron pedidos a la API de LocationIQ y el resultado fue guardado en una instancia de almacenamiento S3 de AWS.

3. Cómputo de $max(f(d))$

Para la conformación del cómputo se utilizó el módulo de cálculo vectorial de Numpy.

VI.4. Procesamiento de imagen en actas de nacimiento

De modo complementario se realizó el procesamiento de actas de nacimiento mediante algoritmos OCR para determinar de manera eficiente las relaciones padres - hijo de los recién nacidos con actas de nacimiento cargadas en CIDI.

A continuación, se describe dicho proceso:

1. Las actas de nacimiento tienen formato pdf, se las transforma a formato jpg utilizando el módulo pdf2image.
2. Se utiliza el módulo pytesseract archivo jpg en un objeta de la clase string. Aquí entra el algoritmo OCR (optical character recognition).
3. Se utilizan expresiones regulares para la obtención de los números de documento que figuran en ese objeto.
4. Se realiza un filtrado para descartar los repetidos.
5. Finalmente se identifica cuál es el documento del nacido y de los padres, exportando el resultado en un objeto de la clase Pandas DataFrame.

VII. Evaluar el grado de confiabilidad de los grupos familiares conformados bajo distintas alternativas

Cuando buscamos computar una métrica de confiabilidad, nos estamos refiriendo a la distancia que nuestra estimación tiene con respecto a valores que se sabe tienen carácter de verdad. Es decir, que representan a la realidad de manera fehaciente.

Teniendo en cuenta esta aseveración, si consideramos como verdad la Encuesta de Bienestar de la Dirección General de Estadísticas y Censos (base real de ahora en adelante) con la cual la presente secretaria cuenta. Podemos medir la distancia de nuestras direcciones computadas de acuerdo al método descripto con anterioridad como la media de diferencias entre el domicilio base o real y nuestro domicilio propuesto para todos los individuos que se tiene registro. En particular, dado que para el propósito que nos compete las direcciones representan una cualidad de la persona, la diferencia es una variable binaria 1 (mismo domicilio en ambos casos) o 0 (domicilio diferente).

Por lo tanto, nuestra métrica de confiabilidad sería:

$$\frac{\sum f(d_{i,r} - d_{i,p})}{n}$$

Donde,

di,r es un vector 2x2[8] y representa la domicilio en nuestra base real en coordenadas.

di,p es un vector 2x2 y representa nuestro domicilio obtenido en coordenadas.

[8] *d = (lat , lon)*

Y por lo tanto nuestra f nos queda,

$$f(d_{i,r} - d_{i,p}) = \begin{cases} 1, & \text{si } d_{i,r} - d_{i,p} < \epsilon \\ 0, & \text{En otro caso} \end{cases}$$

Es decir, 1 en caso de que estén lo suficientemente cerca, definido con cierto grado de aproximación, por ejemplo 0.000001, y 0 en caso de que ambas direcciones se encuentren más alejadas.

Una vez computada esta distancia entre nuestra base real y predicha. La actualización informativa implica realizar una correspondencia entre los identificadores de personas actuales y su nuevo domicilio. Es decir, vamos a tener una relación $personas$ 1:1 $domicilio$s con los nuevos domicilios obtenidos.

Este criterio permitirá una mejor evaluación del grado de confiabilidad de los grupos conformados a los fines de que puedan extraerse conclusiones que permitan plantear nuevos paradigmas de estudio de la temática, especialmente en aquellos casos en que el ciudadano esté en proceso de evaluación para acceder a un programa provincial que incluya, entre las condiciones de elegibilidad, variables mensurables a nivel de grupo familiar.

VIII. Adaptaciones en la delimitación de grupo familiar

En función de las herramientas desarrolladas en los apartados previos, respecto a la dimensión de análisis de conformación del grupo familiar en la PBS, se propuso el desafío de conformar y testear la calidad / confiabilidad de delimitaciones alternativas de la variable de grupo familiar.

En set sentido, luego de haber transitado el proceso de depuración de duplicados en la PBS se trabajó sobre la normalización de grupos familiares, teniendo en cuenta las múltiples fuentes de información disponibles y la información relativa a domicilios, que también podría contribuir a la delimitación de un grupo “conviviente”.

En este sentido, la primera regla definida fue que, si la persona analizada en la PBS ya disponía de un grupo constituido bajo el campo de grupo único, el mismo se mantenía. Por su parte, si la persona no tenía grupo único, pero sí tenía grupo familiar se le asignaba esa codificación a todos sus integrantes.

Por su parte, como tercer subconjunto de personas, se analizó aquellos casos que no cumplían ninguna de las dos condiciones anteriores, a las cuales se les generó un grupo familiar “ficticio” con codificación específica.

Además, se efectuó una clasificación de las personas pertenecientes a grupos familiares de beneficiarios de PAICor que no cruzaron con la base de PBS, contrastando las características de su grupo familiar de PAICor, respecto de grupo único. Frente a esto, pueden darse 6 casos, los cuales son descriptos abajo y en función a ello se definió como proceder, construyendo la variable “grupo2” y “calidad grupo2”.

Tabla Nª 14: Evaluación de adaptaciones en grupo familiar (parte 1)

Casos posibles de la union de bases PAICOR y Base Madre		
	Caso	**Calidad_grupo2**
No esta en BUC y todos sus familiares tampoco	**2**	**2**
Cantidad de personas en caso 2	19,347	
Grupos familiares en caso 2	6,410	
Grupos familiares unipersonales en caso 2	977	
No está en BUC y algunos de sus familiares no están en BUC	**3**	**3**
Cantidad de personas en caso 3	15,230	
Grupos familiares en caso 3	6,187	
Grupos familiares unipersonales en caso 3 (deberia ser cero)	0	
No está en BUC pero toda su familia si y figura con mismo grupo	**4**	**2**
Cantidad de personas en caso 4	11,540	
Mismo id grupo de grupo unico del resto de su familia	11,540	
No está en BUC pero toda su familia si y figuran separados en BUC	**5**	**3**
Cantidad de personas en caso 5	62,866	
Grupos familiares en caso 5	17,828	
Grupos familiares unipersonales en caso 5	9	
No está en BUC y no tiene id_grupo paicor	**6**	4
Cantidad de personas en caso 6	0	

Frente a esto, se creó la variable “grupo 3”, definida como grupo 2 más los integrantes de grupos familiares de beneficiarios del programa Salas Cuna, los cuales fueron reasignados con el mismo criterio que a los integrantes de grupos familiares PAICor en esta misma situación.

En cuanto a los 100 casos que no cruzaron con la base PBS, se realizó una asignación de grupo familiar según el caso donde clasificaba, obteniéndose los resultados expuestos en la Tabla 15. De esta manera, se generó una variable de grupo familiar alternativa, definida como “grupo_adaptado”. Sobre la misma, se realizaron algunas correcciones de registros sin pkid completo.

Posteriormente, esta variable se actualizó renombrándola como “grupo_adaptado2” realizando una actualización que permitió asociar a niños incorporados en la base PBS (de acuerdo a registros de nacimiento provistos por Registro Civil) de 0 a 4 años de edad.

Tabla Nª 15: Evaluación de adaptaciones en grupo familiar (parte 2)

Casos posibles de la union de bases PAICOR y Base Madre	Caso	Calidad_grupo
No esta en BUC y todos sus familiares tampoco	**2**	**2**
Cantidad de personas en caso 2	3	
Grupos familiares en caso 2	2	
Grupos familiares unipersonales en caso 2	1	
No está en BUC y algunos de sus familiares no están en BUC	**3**	**3**
Cantidad de personas en caso 3	7	
No está en BUC pero toda su familia si y figura con mismo grupo	**4**	**2**
Cantidad de personas en caso 4	8	
No está en BUC pero toda su familia si y figuran separados en BUC	**5**	**3**
Cantidad de personas en caso 5	82	
No está en BUC y no tiene id_grupo paicor	**6**	**4**
Cantidad de personas en caso 6	0	

Estos nuevos integrantes de la PBS se incorporaron sin grupo familiar anexado, por lo que se utilizó grupo único para asociarlos al grupo familiar de mejor correspondencia.

Por otro lado, también se presentó la dificultad de encontrar niños a los que no pudo asignársele un grupo familiar específico de acuerdo a la información disponible, de manera que se les generó un grupo familiar de fantasía para demarcar que en esos casos resta trabajo por realizar en materia de asignación de grupo.

Por último, se realizaron otras mejoras y test de calidad de los grupos familiares alternativos que se construyeron, incorporando en el análisis la dimensión de la marca temporal de actualización de la información del domicilio y el grupo familiar, de manera que pudieron delimitarse diferentes niveles de confiabilidad de los grupos específicos, incluyendo en esta definición también a valor indicado por la marca temporal del dato.

Al respecto, cabe destacar que los progresos realizados en materia de utilización del Formulario Único de Postulantes (FUP) para el acceso a programas provinciales, con el requerimiento de que los postulantes actualicen su delimitación de grupo familiar, sumado a la integración de fuentes de grupos familiares en grupo único, contribuyeron a que la definición específica y la calidad de los grupos familiares conformados mejore significativamente.

IX. Conclusiones y recomendaciones

Una vez concluido el presente estudio es posible trazar algunas conclusiones fundamentales de los logros que pudieron alcanzarse con el mismo y los próximos pasos que podrían transitarse.

En primer lugar, y como fue destacado en la segunda sección del documento, se destaca la practicidad y utilidad de contar con una base unificada de información socio-demográfica de la población. El esfuerzo realizado para lograr esta compilación, resolviendo inconvenientes ligados a la aparición de casos duplicados en base de datos, permitió no sólo facilitar las instancias de procesamiento y análisis de esta información sino también aumentar su nivel de confiabilidad.

Asímismo, sirvió para evaluar y reducir el grado de discrepancias entre las bases integrantes, ya sea por incompatibilidades en la información incluida en estos registros como también por faltante de datos.

En este sentido, pudieron desarrollarse valiosas soluciones a la presentación de datos duplicados, bajo distintos enfoques que permitieron abordar y resolver en gran medida la ocurrencia de esta problemática. Los ejemplos implementados a lo largo de este proyecto, sirvieron para que los cuadros técnicos de gobierno pudieran aplicarlos a escala con muy alentadores resultados en materia de identificación unívoca de personas en la base, comentados en el quinto título del trabajo.

Por otra parte, de acuerdo a lo desarrollado en la tercera sección, se estudiaron maneras alternativas de delimitación de grupos familiares, partiendo de la base de que existe una sobre-representación de grupos unipersonales y la importancia de lograr delimitar correctamente la realidad de estos hogares, puesto que muchos programas provinciales se basan en un análisis de la realidad socio-económica del grupo familiar como condición de admisibilidad.

En este sentido, se presentó un diagnóstico relativo a la evolución observada en la definición del Grupo Familiar de acuerdo a sus distintos orígenes hasta llegar a la actual concepción del “Grupo único”.

Posteriormente, en la cuarta sección, se desarrolló un ejemplo demostrando cómo es posible enriquecer las bases administrativas con un paradigma de georreferenciación de datos en términos de latitud y longitud, lo cual abre la puerta a la posibilidad de realizar análisis de distribución geo-espacial de características socio-demográficas de los habitantes y de la percepción / acceso a programas provinciales en vigencia.

En el mismo sentido, se valora y recomienda la compilación de datos geo-referenciados en serie de tiempo, a los fines de poder observar las dinámicas geográficas de indicadores sociales, percepción / dependencia de programas, etc.

Asimismo, en las últimas secciones (sexta a octava) se implementaron metodologías de trabajo acordes a las nuevas herramientas disponibles para la explotación y la ciencia de datos basada en el objetivo de efectuar ejercicios que permitan testear la calidad de conformación de grupos familiares alternativos, simulando diferentes tipos de configuraciones y pudiendo evaluar objetivamente las conformaciones encontradas.

En particular, se desarrolló e implementó una metodología de agrupación en base a domicilios alternativos, de acuerdo a información provista por el Sistema de Identificación Nacional Tributario y Social (Sintys) y fuentes de información del Gobierno Provincial.

En base a estos insumos, se realizaron ejercicios alternativos en los que se dimensionaba, por ejemplo, el domicilio más actual y el domicilio con mayor cantidad de repeticiones entre fuentes alternativas, entre otros. Esta propuesta sirvió para definir al domicilio habitual del individuo y ensayar cómo quedaría conformado su grupo familiar frente a estas variantes.

El resultado de estas pruebas y la transmisión a los cuadros técnicos de Gobierno de los criterios de tratamiento de datos aplicados servirá para una mejor y permanente evaluación de los grupos familiares probables para cada persona incluida en la base de datos, especialmente si este criterio es tomado en consideración para la delimitación del acceso / no acceso a un beneficio social conferido por la provincia.

En suma, los avances producidos en este proyecto y los intercambios permanentes realizados con los cuadros técnicos del Gobierno Provincial permitieron la

transferencia de herramientas objetivas y conocimiento específico para el uso de nuevas técnicas y métodos de la ciencias de datos y machine learning para el tratamiento y análisis de registros administrativos. En este sentido, se logró enriquecer y dar mayor eficiencia a los criterios de explotación de datos.

Por su parte, los desarrollos realizados contemplaron además la posibilidad de que la explotación de datos pudiera ser presentada y dispuesta a sus usuarios finales en formato de tableros de datos, específicamente en el marco del trabajo con el software Microstrategy.

En este sentido, se tomaron en consideración los requerimientos técnicos en los datos que utiliza como insumo este programa para la posterior elaboración de visualizaciones y la incorporación de botones interactivos que permitan al usuario de la información trabajar con flexibilidad en sus análisis, con herramientas gráficas y tabulados de interés que resumen la información y facilitan su interpretación.

Al respecto, a lo largo del plazo de ejecución del proyecto se produjeron intercambios con los equipos técnicos del Gobierno Provincial encargados de la disposición de información proveniente de la PBS en tableros de Microstrategy, lo cual resultó en un intercambio altamente productivo para el logro de los objetivos propuestos en el trabajo.

La perspectiva con la que fue abordada el trabajo, centrada en la transferencia a cuadros técnicos y el cumplimiento de los objetivos trazados al inicio, permitió un enriquecimiento mutuo y una nueva ventana de oportunidades para que la gestión pública aproveche las ganancias acumuladas en la capacidad de procesamiento de datos administrativos y confíe en estos insumos para los procesos de toma de decisiones, a la vez que continúe apoyando el desarrollo de proyectos de este tipo.

En igual sentido, cabe destacar la flexibilidad con la que pudo trabajarse a lo largo de todo el plazo de ejecución del proyecto, periodo en el cual sin desviarse de los objetivos trazados, que constituían el desarrollo estructural deseado por el Gobierno Provincial, se pudieron realizar adaptaciones en el foco de análisis y la resolución de situaciones problemáticas concretas que se presentaron como ejes de interés en la realidad coyuntural tan particular en que se desarrolló el proyecto.

En todo momento, se buscó complementar y mejorar el trabajo realizado con anterioridad a la ejecución del proyecto, así como también adecuando los procedimientos de ciencia de datos recomendados a las oportunidades específicas del Gobierno Provincial, así como también a los requerimientos de seguridad informática correspondientes al caso.

Sin lugar a dudas, analizando en retrospectiva al punto de inicio en la ejecución de este proyecto, puede destacarse que el Gobierno Provincial ha ganado en su capacidad técnica y ornamental de programación para la ejecución de rutinas de interés, previendo que éstas puedan ser sostenidas y perfeccionadas en el futuro.

Referencias

Documentación de módulos infórmaticos utilizados:

- Pandas https://pandas.pydata.org/
- Numpy https://numpy.org/
- Itertools https://docs.python.org/3/library/itertools.html
- Pytesseract https://pypi.org/project/pytesseract/
- Pdf2image https://pypi.org/project/pdf2image/
- OCR teseeract https://github.com/UB-Mannheim/tesseract/wiki
- Engine https://docs.sqlalchemy.org/en/14/core/engines.html
- Cx_oracle https://oracle.github.io/python-cx_Oracle/
- Ipystata https://github.com/TiesdeKok/ipystata
- Vaex https://vaex.io/docs/index.html

Bibliografía:

Wes McKinney (2017). "*Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*", documento de trabajo, Estados Unidos de América.

Anexo: Ejemplos del problema de duplicados

Para estos ejercicios, vamos a suponer dos bases de datos “Padrón electoral” y “Tarjeta Social”:[9]

Ejemplo 1: Caso Ideal

Supongamos que tenemos el caso:

Padrón electoral

	id_sexo	pai_cod_pais	nro_documento	id_numero
0	1	ARG	3039624	1243
1	2	ARG	3039625	2486
2	1	BOL	3039624	1243

Tarjeta social

Resultante

	id_sexo	pai_cod_pais	nro_documento	id_numero
0	1	ARG	3039624	1243
1	2	ARG	3039625	2486
2	1	BOL	3039624	1243
2	1	BOL	3039624	1243

En este ejemplo podemos evidenciar el caso en el que dos bases de comparten la misma información para todas las variables, por ende, su cruce es exitoso y la identificación es unívoca.

[9] Ver ejemplo 3 en repositorio.

Ejemplo 2: Problema de datos faltantes

Supongamos esta otra alternativa:

Padrón electoral

	id_sexo	pai_cod_pais	nro_documento	id_numero
0	1	ARG	3039623	1243
1	2		3039624	2486
2	1	BOL	3039625	3729

Tarjeta social

	id_sexo	pai_cod_pais	nro_documento	id_numero
0	1	ARG	3039623	1243
1	2		3039624	2486
2	1		3039625	3729

Resultante

Aquí vemos que ambas bases tienen valores faltantes en la variable pai_cod_pais. En estos casos, se pueden dar dos alternativas:

1. El “missing” coincide, entonces el cruce es exitoso porque se interpreta que es la misma persona

	id_sexo	pai_cod_pais	nro_documento	id_numero
0	1	ARG	3039623	1243
1	2		3039624	2486
2	1	BOL	3039625	3729
3	1		3039625	3729

2. En una base hay “missing” y en la otra no, por lo tanto se interpreta que son personas distintas y se genera un duplicado.

Ejemplo 3

Problema de variable faltante

Padrón electoral

	id_sexo	nro_documento	id_numero
0	1	3039624	1243
1	2	3039625	2486
2	1	3039624	1243

Tarjeta social

	id_sexo	nro_documento	id_numero
0	1	3039624	1243
1	2	3039625	2486
2	1	3039624	1243

Resultante

	id_sexo	nro_documento	id_numero
0	1	3039624	1243
1	1	3039624	1243
2	1	3039624	1243
3	1	3039624	1243
4	2	3039625	2486

En este ejemplo nos encontramos con dos bases iguales en las cuales debido a la omisión de una variable tenemos un duplicado en cada una. De esta forma cuando se realiza la intersección sin relación de preferencia, el resultado es el producto externo entre ambos vectores. Y consecuentemente, la cantidad de duplicados es el doble.