TP1: Manejo y visualización de datos

Laboratorio de Datos - 1° cuatrimestre 2023

Grupo: las_yararas
Altamirano Ailen, Rio Francisco, Ruz Veloso Luciano

Resumen

En este trabajo, analizamos datos abiertos correspondientes al Padrón de Operadores Orgánicos Certificados de la República Argentina y a la mediana del salario bruto de los trabajadores registrados del sector privado. El objetivo del análisis fue caracterizar la relación entre el desarrollo de la actividad según la provincia y el salario percibido por los trabajadores del sector orgánico. Para ello, se incorporaron tres fuentes de datos secundarias, y se normalizaron todos los *datasets*. Adicionalmente, se realizó un análisis de la calidad de los datos siguiendo la metodología GQM, y a partir de las métricas resultantes se tomaron decisiones con respecto a la limpieza de los datos. A partir de gráficos y de consultas de SQL, se determinó que la actividad 'Agricultura, ganadería, caza y servicios relacionados' está asociada a salarios menores que el resto de las actividades, especialmente en las provincias que poseen mayor cantidad de operadores orgánicos. Esto es consistente con que las actividades de la industria primaria requieren mayor mano de obra, lo que podría dar lugar a una oferta laboral más alta en relación a la demanda, lo que a su vez podría provocar la disminución de los salarios.

Introducción

La disponibilidad de datos abiertos se incrementa día a día, lo cual ha llevado a una creciente demanda por parte de distintos sectores en la exploración relaciones entre diferentes variables. En este contexto, el gobierno nacional argentino creó el portal *datos.gob.ar,* el cual permite el acceso a información pública y datos gubernamentales. En dicha página es posible encontrar datos de diversas áreas, como salud, educación, seguridad, medio ambiente, economía, entre otros.

En este trabajo, nos centramos en los datos correspondientes al Padrón de Operadores Orgánicos Certificados de la República Argentina. Este padrón es un registro nacional de los productores y operadores de la cadena de producción orgánica que fueron certificados por organismos reconocidos por el Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA). Un operador orgánico certificado es aquel operador que produce utilizando técnicas y prácticas de cultivo que cumplen ciertos requisitos y estándares. Dichos requerimientos se basan en el uso de métodos naturales y sostenibles que respetan el medio ambiente, la biodiversidad y la salud de los consumidores, evitando el uso de químicos sintéticos, el uso de agroquímicos de síntesis química (herbicidas, fertilizantes, pesticidas) y el uso de organismos genéticamente modificados (GMOs). El objetivo de la certificación es garantizar la calidad y la seguridad de los productos orgánicos y ofrecer a

los consumidores una alternativa más saludable y sostenible. Pero estos requerimientos que recaen sobre los operadores para ser certificados provocan que se necesite mayor mano de obra para garantizar su cumplimiento. Por ello, a partir de este *set* de datos, junto con una fuente de datos que contiene la mediana del salario bruto de los trabajadores registrados del sector privado, nos proponemos estudiar la relación entre el desarrollo de la actividad según la provincia y el salario percibido por los trabajadores del sector orgánico.

Para cumplir dicho objetivo, incorporamos tres fuentes de datos secundarias: un listado de localidades, un diccionario de departamentos y un diccionario de clases. Dichas fuentes secundarias permitieron vincular los dos set de datos principales. Para vincular los *sets* de datos de manera correcta, fue necesario normalizar todas las tablas, analizar la calidad de los datos y realizar una limpieza de los mismos.

Decisiones tomadas

Para garantizar la integridad y consistencia de los datos, se tomaron las siguientes decisiones:

- Se agregó el atributo "id_registro" como clave de fantasía en la tabla padrón. Una clave alternativa sería la combinación de atributos {establecimiento, razon_social}. Se decidió utilizar una clave de fantasía ya que el atributo 'establecimiento' contiene NULLS, lo cual violaría la integridad de la clave primaria.
- Se consideró que los atributos "clae2_desc" y "letra_desc" en la tabla de Clases eran categorías fijas y se trataron como atributos atómicos.
- Se omitió la columna "pais" y "pais_id" ya que todos los registros corresponden a Argentina.
- La columna "localidad" de la tabla padrón no se tomó en cuenta debido a la gran cantidad de valores nulos presentes.
- Se eliminaron los acentos y se convirtieron todos los nombres de provincias, localidades y departamentos a mayúsculas. Esto fue realizado para lograr que los distintos *set* de datos sean consistentes entre sí y facilitar las comparaciones entre ellos.
- En la tabla "localidades", se modificó el nombre de la provincia "TIERRA DEL FUEGO, ANTARTIDA E ISLAS DEL ATLANTICO SUR" a "TIERRA DEL FUEGO" para lograr consistencia en los datos entre las tablas.
- En la tabla "padron", se modificó el nombre de "CIUDAD AUTONOMA BUENOS AIRES" a "CIUDAD AUTONOMA DE BUENOS AIRES" para lograr consistencia en los datos entre las tablas.
- En la columna "departamentos" de la tabla "padrón", se asignó el código de departamento correspondiente utilizando la tabla de "localidades censales" como referencia. En los casos en que la columna "departamento" representaba una localidad, se asignó el número de departamento al que pertenecía dicha localidad. Los operadores que no se le pudo identificar el departamento fueron eliminados.
- Se eliminaron los registros correspondientes a los operadores orgánicos de "LOS CARDALES" en la provincia de Buenos Aires, ya que pertenecían a diferentes departamentos y no se pudo determinar a cuál correspondían.

- Se extrajo manualmente el código correspondiente al departamento "CIUDAD AUTONOMA DE BUENOS AIRES" del diccionario de departamentos y se asignó a todos los registros correspondientes a dicha localidad. Esto porque la tabla “localidades” no tenía un código asignado.
- Se corrigió el error de tener dos códigos diferentes para el departamento "Ushuaia" en el diccionario de departamentos y la tabla de localidades, utilizando la información obtenida de búsquedas en Internet.
- Se ajustaron los nombres de departamentos en la tabla correspondiente para que coincidieran con los de la tabla correspondiente a localidades.
- Se eliminaron los registros de padrón cuyo rubro era ‘SIN DEFINIR’, dado que la información del rubro es lo que permite definir la actividad correspondiente. Conocer la actividad de cada operador es indispensable para el objetivo de este trabajo.
- No se consideró rubro como un atributo atómico, dado que un operador puede pertenecer a más de un rubro. Por ejemplo, un operador orgánico que cosecha uvas y produce vinos tendría que ser clasificado posteriormente como agricultor y también como productor de bebidas.

Estas decisiones fueron tomadas con el objetivo de garantizar la calidad y coherencia de los datos para su posterior análisis e interpretación. En la sección “Procesamiento de Datos” se brindan más detalles sobre los problemas de calidad encontrados y su magnitud.

Procesamiento de Datos

Consideramos como operador orgánico a cada combinación entre razón social y establecimiento, así registros coincidentes en su razón social pero cuyo establecimiento difiere serían considerados como operadores distintos. En relación a esto, se generó el atributo “ID_registro” en la tabla padrón, el cual es un identificador único de registro correspondiente a Operadores Orgánicos y será utilizado como clave primaria. Una clave alternativa sería la combinación de atributos {establecimiento, razon_social}. Se decidió utilizar una clave de fantasía ya que el atributo ‘establecimiento’ contiene NULLS, lo cual violaría la integridad de la clave primaria.

Se adaptaron las tablas para que se encuentren en primera forma normal. La tabla padrón de operadores orgánicos certificados poseía dos atributos no atómicos: “rubro” y “productos”. En ambos casos creamos una tabla separada que vincula el id_registro de cada operador con sus respectivos rubros y productos.

Se identificó el conjunto minimal de las dependencias funcionales con el objetivo de llevar todas las tablas a tercera forma normal, es decir que no existan atributos que dependan parcialmente de una clave primaria ni dependencias transitivas.Las dependencias funcionales identificadas a partir de la semántica de los atributos fueron:

id_registro → {codigo_departamento, categoria_id, certificadora_id, razon_social, establecimiento, rubro}

id_provincia → provincia

categoria_id → categoria_desc

Certificadora_id → certificadora_desc

rubro → clase_id

clase_id → {clase_desc, letra}

letra →letra_desc

codigo_departamento → {nombre_departamento, id_provincia}

Localidad.id → {categoria, centroide_lat, centroide_long, id_municipio, nombre}

{centroide_lat, centroide_long} → {Localidad.id}

municipio_id → {municipio_nombre, codigo_departamento}

{fecha, clae2, codigo_departamento} → w_median

Se plantea entonces el siguiente Diagrama Entidad Relación (DER) (Figura 1) y el modelo relacional (Figura 2) que se obtuvo a partir del mismo.

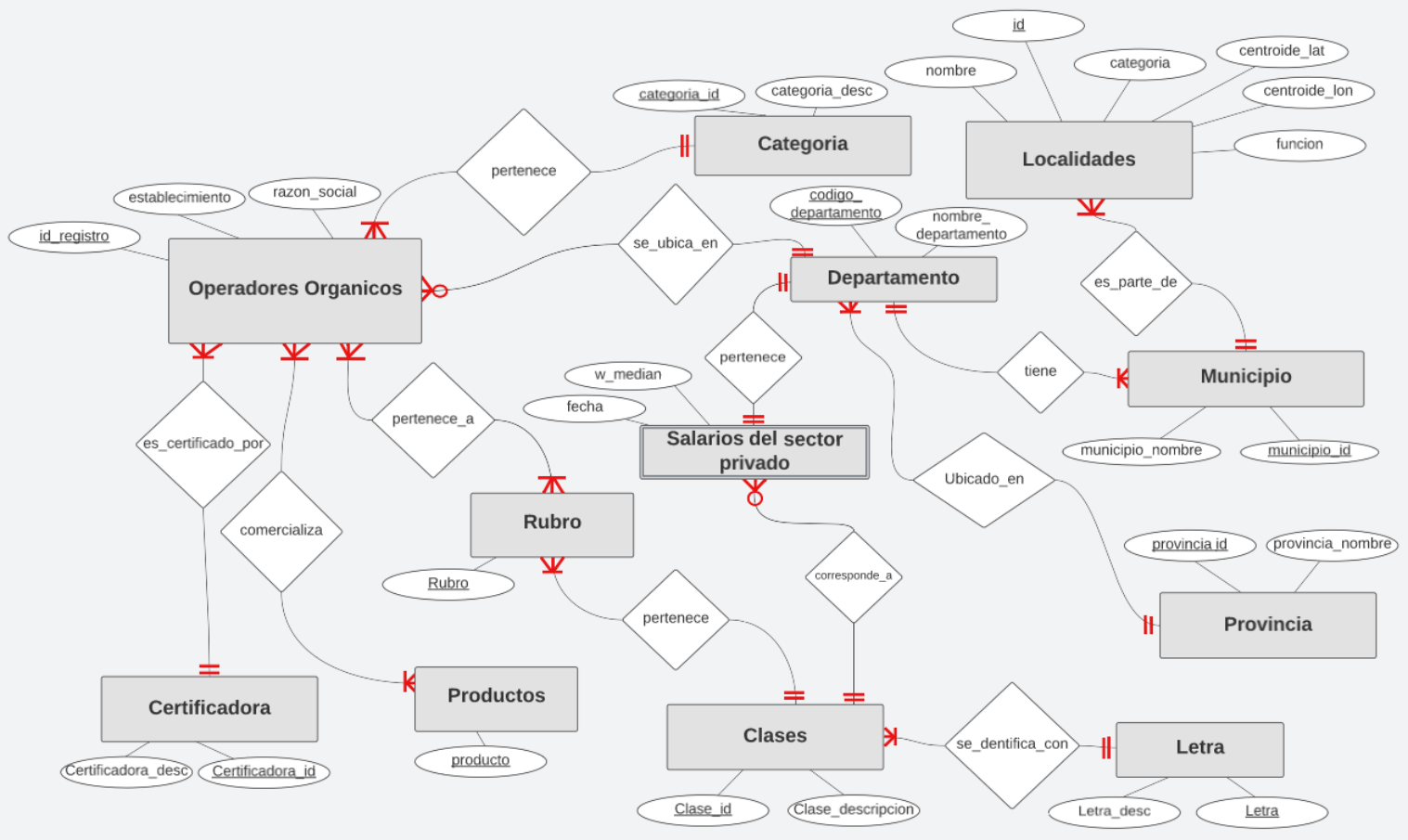


Figura 1. Diagrama Entidad Relación (DER)

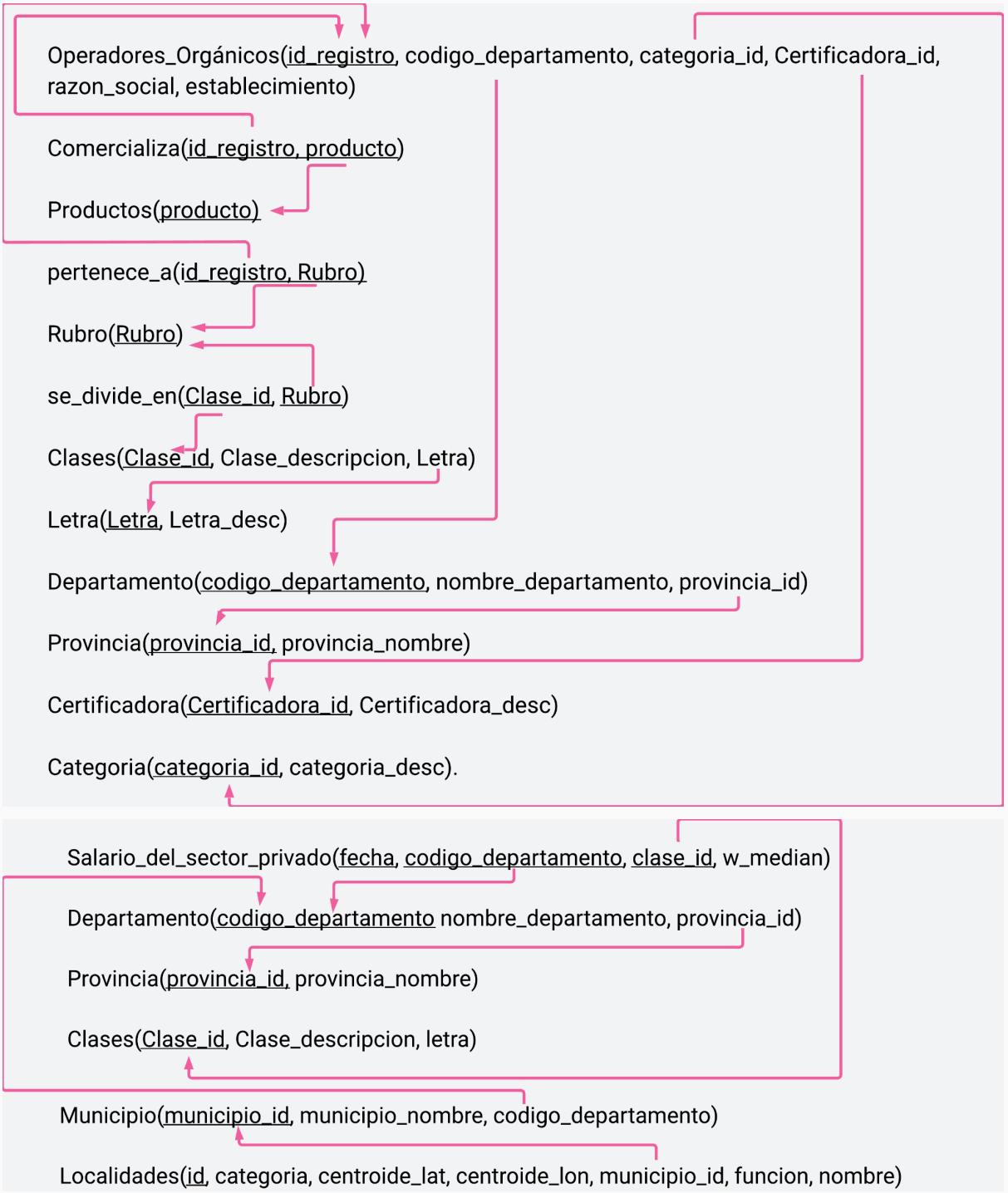


Figura 2. Modelo relacional. Las flechas apuntan desde una clave foránea a la clave primaria correspondiente. Se decidió representar el Modelo en imágenes separadas para facilitar su lectura, representando algunas relaciones en ambas imágenes.

En el caso de Operadores Orgánicos, para lograr que quede en primera forma normal (1FN), se crearon las tablas de 'Productos' y 'Rubro' dejando así valores atómicos en cada una de las tablas. Por otro lado, para conseguir que se cumpla la tercera forma normal (3FN) se crearon tablas para provincia, departamento, categoría y certificadora, dejando en Operadores sólo el id correspondiente como *primary key* para poder relacionar las tablas (menos en el caso de provincia ya que con el código departamento se puede saber a qué

provincia pertenece). Asimismo, 'Departamento' tiene como *primary key* 'codigo_departamento'. Rubro, por su parte, se relaciona con 'Clases' y esta última con 'Letra', todas con sus respectivas *keys*.

Además, 'Salarios_del_sector_privado' se relaciona con 'Departamento' y 'Clases' conservando las respectivas claves primarias. En este caso, se consideró que 'fecha', 'codigo_departamento' y 'clase_id' eran las *keys* de esta tabla.

En el caso de la tabla 'Localidades' se tiene como *key* el 'id' y se conservó 'municipio_id', descartando así 'departamento_id' y 'provincia_id' ya que con saber el municipio se pueden conocer estos últimos.

El análisis de calidad de los datos fue realizado siguiendo la metodología GQM (*Goal Question Metric*). Algunos aspectos relacionados a la calidad de los datos fueron ya mencionados en la sección **Decisiones tomadas.**

Para lograr que la tabla Operadores Orgánicos contenga la información indicada en el Modelo Relacional (Figura 2) fue necesario asignarle un código de departamento a cada uno de los registros de dicha tabla. Para ello, utilizamos los nombres de departamento incluidos en la tabla padrón antes de ser normalizada. Utilizando solo los nombres de departamento, no fue posible asignarle un código de departamento al 74,19% de los registros. Por ello, decidimos también utilizar los nombres de localidades presentes en la tabla de localidades censales, asignando el código de departamento al que pertenece cada localidad a los registros del padrón de operadores. Al momento de realizar esto, el porcentaje de los departamentos no asociados a ningún código disminuyó a 8,39%. Decidimos eliminar los datos correspondientes a estos operadores, conservando así aproximadamente el 91% de los registros. Para que la asignación de códigos de departamentos fuera exitosa, los datos correspondientes a los distintos departamentos debían ser consistentes entre tablas. Por un lado los códigos de departamentos deben ser consistentes entre las tablas diccionario de departamentos y localidades censales. Debido a eso nos preguntamos cuántos departamentos de la tabla diccionario de departamentos tienen código distinto al de la tabla correspondiente a localidades censales. Para responder dicha pregunta, calculamos como métrica el cociente entre la cantidad de departamentos con código distinto y el número de departamentos totales de la tabla diccionario de departamentos, obteniendo un valor de 0.0019 (0.19%). A su vez, los departamentos identificados con el mismo código en las tablas diccionario de departamentos y localidades censales deben coincidir en su nombre de departamento. Al calcular la proporción de departamentos de la tabla diccionario de departamentos que coinciden en código pero no en nombre con un departamento de la tabla de localidades censales, obtenemos 0.0176 (1.76%).

Por otro lado, un problema de calidad de datos importante consiste en que la mediana de los salarios no puede ser negativa. El 22,42% de valores de mediana de salario eran menores que cero, por lo tanto poseían un valor que no pertenece al dominio de la mediana de los salarios. Esos valores negativos no son un reflejo de la realidad, sino que al no poder preservar la anonimidad de ciertos trabajadores se les asignó un salario de "-99". Esta práctica fue realizada por el INDEC con el fin de proteger información personal (CEP, 2022, p. 3). Se decidió eliminar estos registros para realizar el resto de los análisis.

Finalmente, surgió un problema cuando se quiso relacionar actividad con salarios, ya que el 8.17% de los rubros estaba sin definir, lo que limitaba el poder realizar este análisis.

Se determinó que al no ser útiles estos datos para cumplir el objetivo del trabajo, fue mejor eliminarlos.

Para vincular los rubros con su clae correspondiente, se generó manualmente la tabla clae_rubro, asignando un número de clae a cada uno de los rubros presentes en la tabla padrón antes de normalizar.

Análisis de Datos

Con los datos resultantes podemos analizar correctamente el desarrollo de estas actividades orgánicas. Para ello se utilizaron consultas de SQL, y las bibliotecas *pandas* y *seaborn*.

Lo primero que se puede observar es que todas las provincias se ven representadas en la tabla de Operadores Orgánicos Certificados. Por el contrario, no pasa lo mismo con los departamentos, ya que hay 311 que no presentan operadores.

Considerando cada rubro individualmente, el que más operadores posee es 'Fruticultura' con 432. La Clae a la que pertenece dicha actividad es 'Agricultura, ganadería, caza y servicios relacionados' cuyo salario promedio, considerando solo los registros correspondientes a diciembre de 2022, fue de ARS$195.856.

La tabla del promedio de los salarios en Argentina (Tabla 1) muestra que el promedio anual, así como su desvío, aumenta año a año. El motivo de esto puede ser la inflación, lo que dificulta comparar los datos a lo largo del tiempo. Al mirar el salario promedio por provincia se observa el mismo fenómeno. Una posible forma de tratar los datos de distintos años para que estos sean comparables consistiría en ajustarlos por la inflación anual o mensual. Para ello, debería utilizarse información del índice inflacionario del INDEC.

Año	Promedio Anual	Desvío Anual
2014	9,847.48	7,045.58
2015	12,921.90	9,392.78
2016	17,139.14	12,382.00
2017	22,226.70	15,877.58
2018	28,172.49	20,640.13
2019	40,686.94	31,269.87
2020	55,805.80	41,081.06
2021	83,412.10	63,349.75
2022	144,996.61	117,350.30

Tabla 1. Promedio y desvío estándar del sueldo por año.

Al mirar la cantidad de operadores, es decir la cantidad de combinaciónes razón social y establecimiento, por provincia (Figura 3), se advierte que Misiones (216), Mendoza (197), Buenos Aires (158) y Río Negro (154) son las que más operadores tienen. Por otra parte, Formosa (2), La Pampa (3), Tierra del Fuego (3), San Luis (5) y Ciudad Autónoma de Buenos Aires (5) son las provincias con menos operadores certificados. Solo consideramos los operadores a los cuales le pudimos identificar su correspondiente departamento.

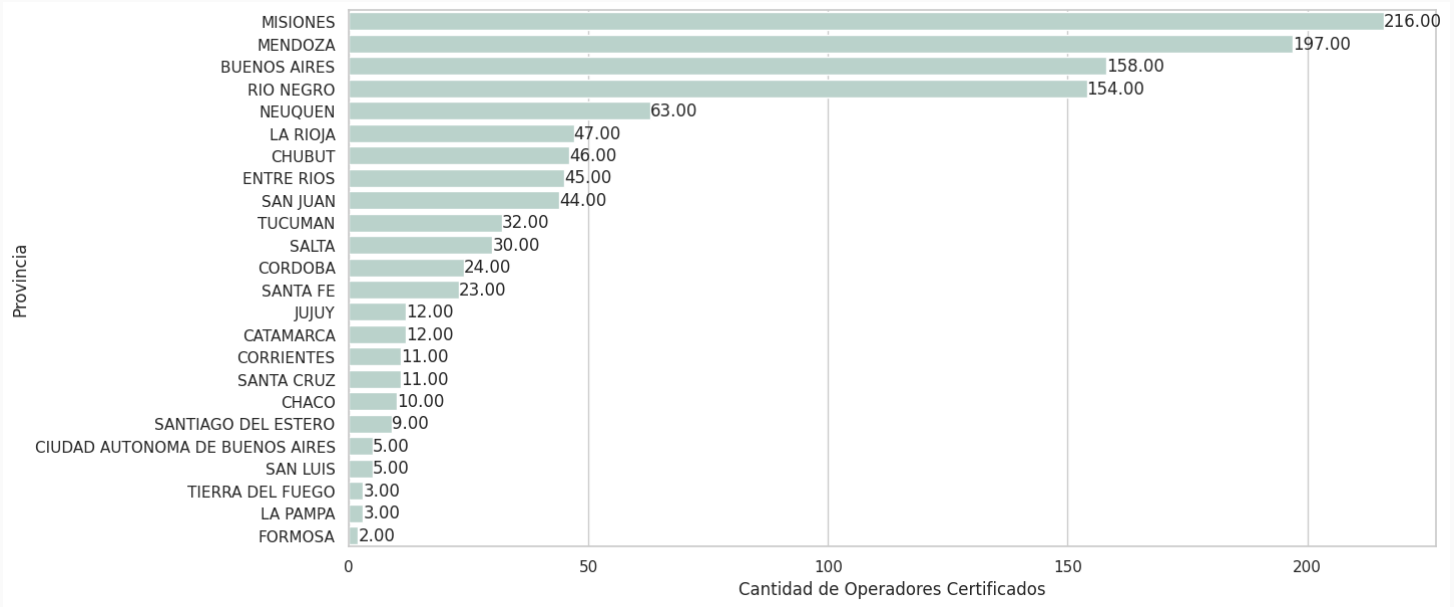


Figura 3. Gráfico de barras representando la cantidad de operadores certificados por provincia.

A su vez, se determinó la cantidad de productos distintos que vende cada operador por provincia (Figura 4). San Luis presenta la mayor mediana de cantidad de productos, seguida por Tierra del Fuego, Chubut y Santa Cruz; mientras que San Juan, Corrientes y Mendoza presentaron las medianas de menor magnitud. A su vez, es posible observar que la dispersión en la cantidad de productos es muy variable según la provincia. Por ejemplo, Neuquen, Santiago del Estero, Rio Negro, Catamarca y Misiones presentan muy poca dispersión en sus datos a comparación del resto de las provincias. También se observan *outliers* o datos atípicos en varias provincias, siendo el de mayor valor correspondiente a Buenos Aires.

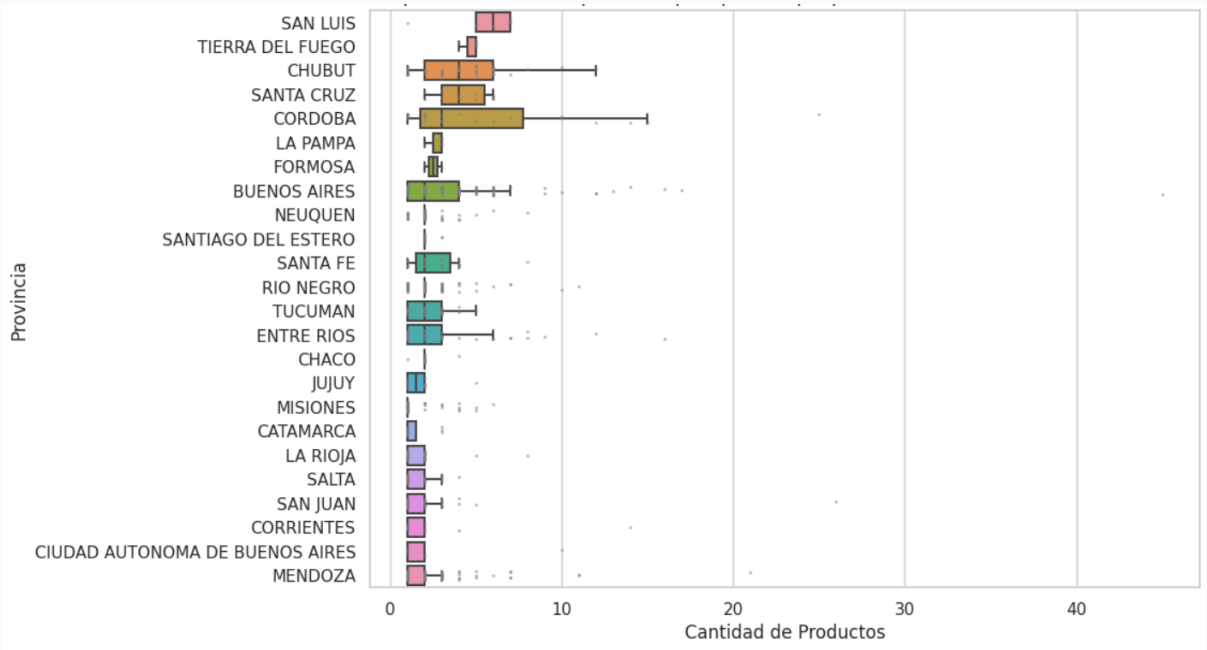


Figura 4. Boxplot correspondiente a la cantidad de productos producidos por operador por provincia. En el gráfico se incluyen los datos de cantidad de productos como puntos grises, junto con el *box* correspondiente a su distribución por provincia. Se consideran datos atípicos a todos aquellos puntos grises que aparecen por fuera de los *'bigotes'* de los *boxplots*.

El siguiente gráfico (Figura 5) busca relacionar la cantidad de operadores con el salario promedio según cada provincia y la actividad realizada (Clae2). Para ello sólo se han utilizado los últimos sueldos del año 2022.

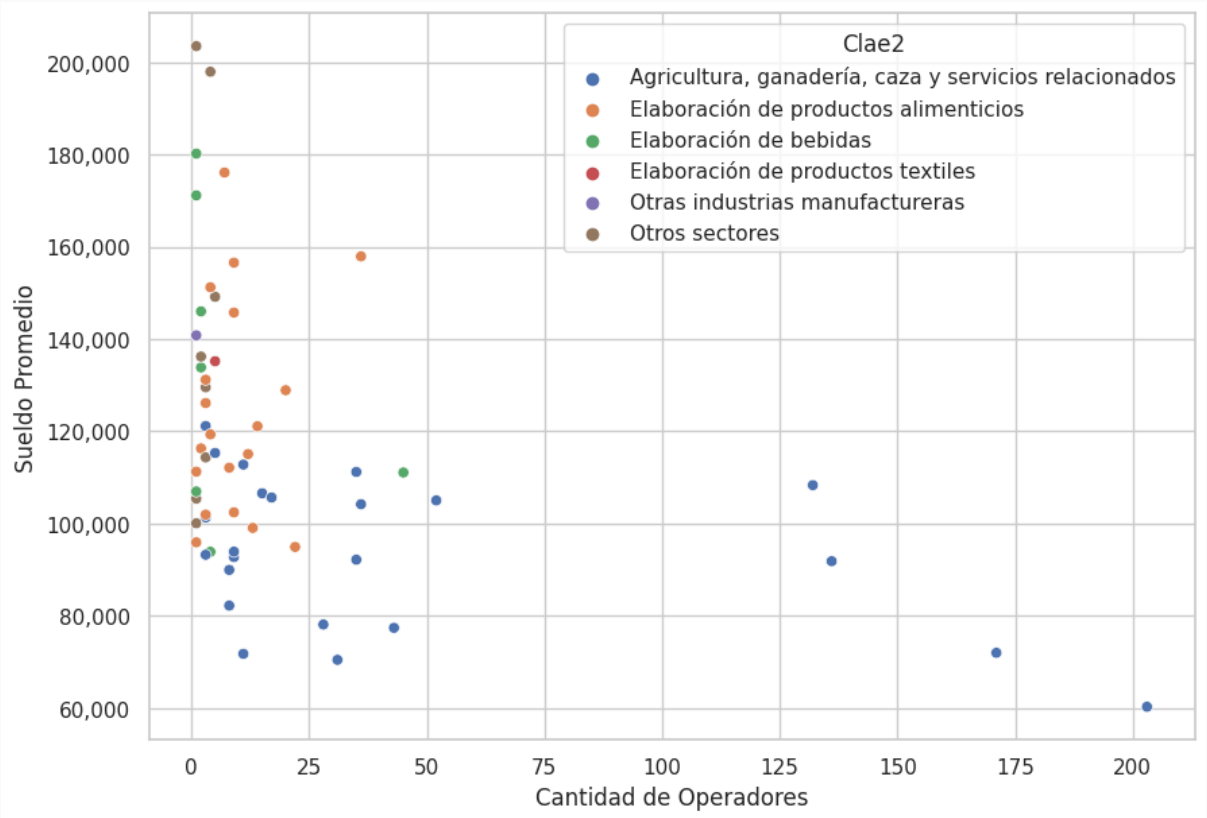


Figura 5. Gráfico de dispersión entre el sueldo promedio y la cantidad de operadores por provincia y por actividad (Clae2). Los colores de los puntos representan su actividad correspondiente. El promedio se calculó en base a los salarios de todo el 2022. Una versión del gráfico coloreando por provincia puede observarse en el Anexo 1.

Se percibe que una menor cantidad de operadores no se relaciona con mejores ingresos, ya que están distribuidos en casi todo el rango de salarios. Por otro lado, a partir de los 125 operadores se ve una disminución del sueldo promedio, pero al ser pocos datos (sólo 4 puntos) no alcanza para sacar conclusiones al respecto. Además, al tener en cuenta la actividad (Clae2), se tiene que 'Agricultura, ganadería, caza y servicios relacionados' es en general la de menor sueldo promedio percibido. En cambio, para la actividad 'Elaboración de productos alimenticios' se observa que la cantidad de operadores no supera los 50, y generalmente los sueldos promedio son mayores que los correspondientes a 'Agricultura, ganadería, caza y servicios relacionados'. El resto de las actividades presenta una baja cantidad operadores y sueldos promedio entre 100.000 y 200.000 ARS aproximadamente. Al repetir el mismo análisis teniendo en cuenta el departamento al que pertenece cada operador (Anexo 2) observamos el mismo patrón: la actividad 'Agricultura, ganadería, caza y servicios relacionados' se asocia a salarios menores y mayor cantidad de operadores que el resto de las actividades.

A continuación, se graficó la distribución de los salarios promedio del año 2022 por provincia (Figura 6) .

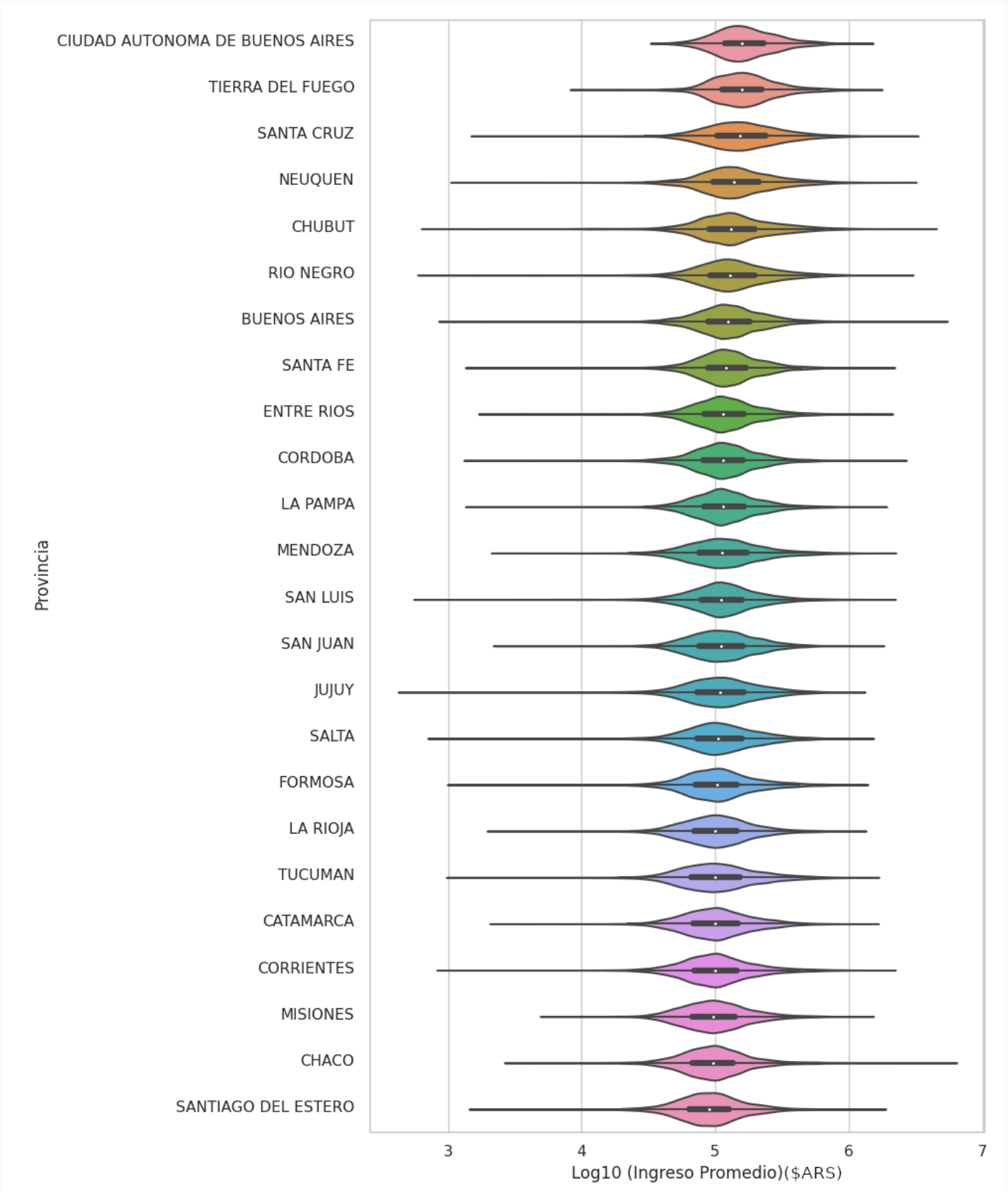


Figura 6. Violín plot correspondiente a los salarios promedio (ARS$) del año 2022 según la provincia. El eje horizontal se representa en escala logarítmica para una mejor visualización de las distribuciones.

En general, se observa que el salario promedio tiene una dispersión similar en las provincias, salvo en el caso de la Ciudad Autónoma de Buenos Aires y Tierra del Fuego que presentan una dispersión considerablemente menor que el resto (Figura 6 A). A su vez, estas dos provincias son las que presentan mayor mediana del salario promedio, seguidas por Santa Cruz y Neuquén.

A partir de esto, decidimos analizar la cantidad de operadores por actividad (Clae2) en las provincias de mayor (Ciudad Autónoma de Buenos Aires y Tierra del fuego) y menor (Chaco y Santiago del Estero) mediana del salario en 2022. Con dicho análisis se obtiene la Figura 7.

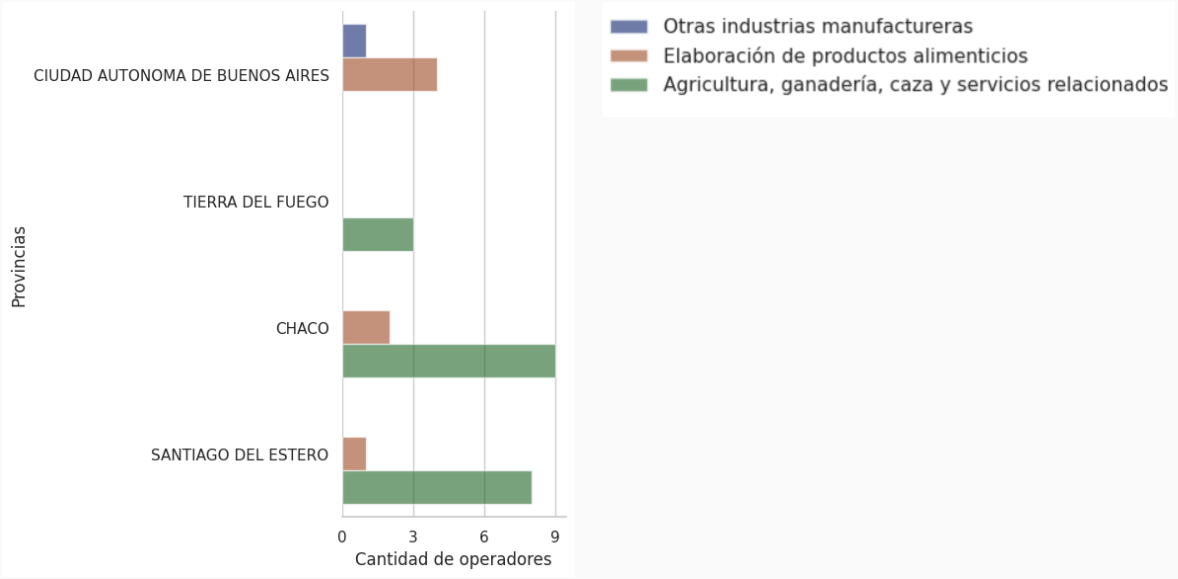


Figura 7. Cantidad de operadores por actividad (Clae2) y por provincia. La actividad se indica con el color de cada barra.

En el caso de la Ciudad Autónoma de Buenos Aires, provincia de mayor salario en promedio en 2022 (Figura 7), es posible observar que no hay operadores correspondientes a la categoría 'Agricultura, ganadería, caza y servicios relacionados'. Los operadores de esta provincia corresponden mayoritariamente a 'Elaboración de productos alimenticios' u 'Otras industrias manufactureras'. Tanto esta provincia como Tierra del Fuego presentan menor cantidad de operadores orgánicos que la mayoría de las provincias restantes (Figura 3).

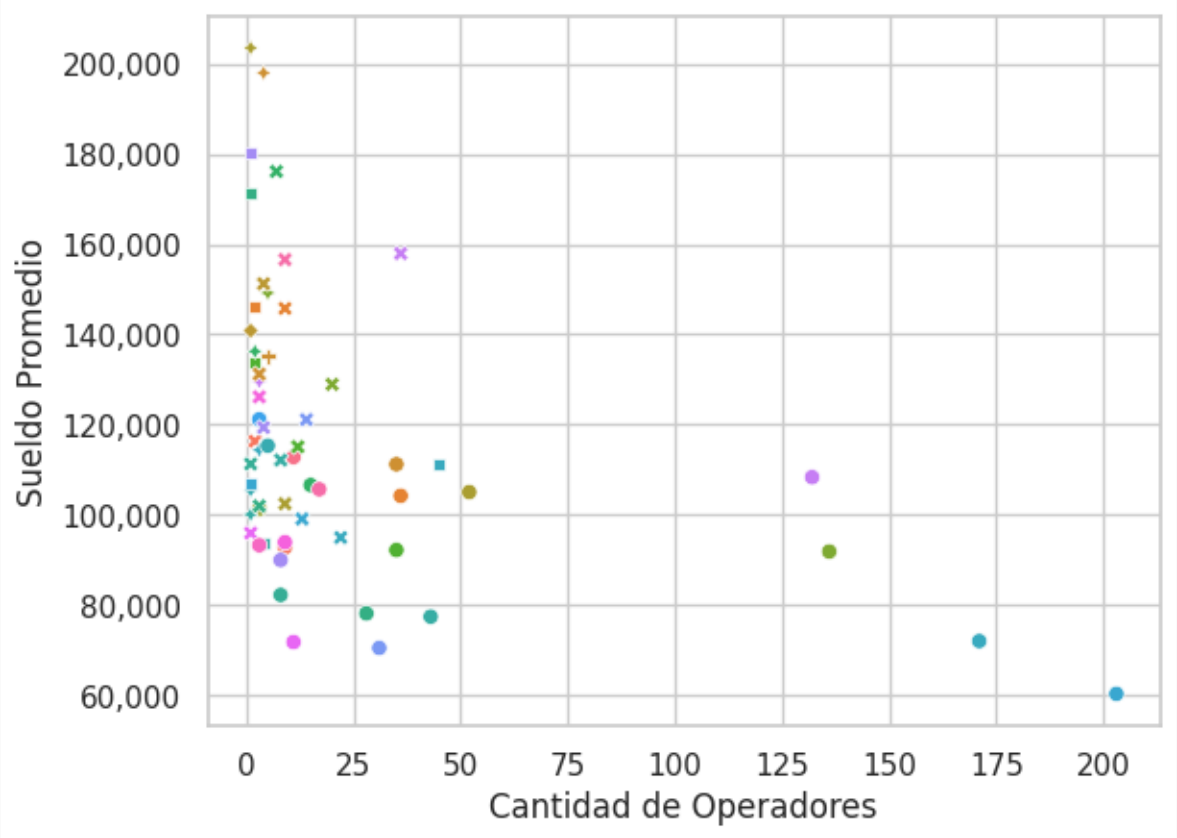
Sin embargo, tanto Tierra del Fuego, como Chaco y Santiago del Estero presentan mayoritariamente operadores categorizados en 'Agricultura, ganadería, caza y servicios relacionados', siendo estas dos últimas las provincias con menor salario en promedio para el año 2022 (Figura 7).

Conclusiones

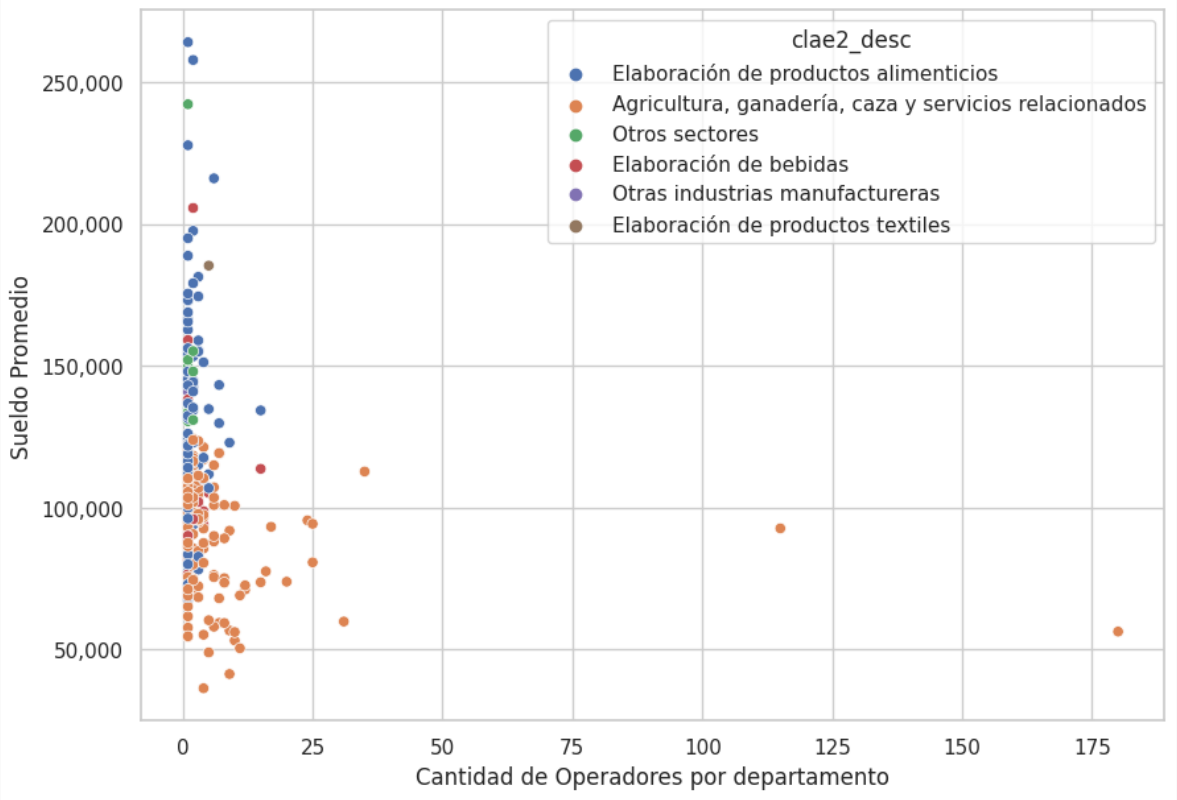
A medida que se procede con las tareas de limpieza y mejora de la calidad de los datos, se evidencia que la base de datos no proporciona suficiente información para llevar a cabo un análisis completo y preciso. Por ejemplo, al momento de querer relacionar operadores orgánicos con salarios, los datos presentan una alta cantidad de valores faltantes o nulos cuando se observa el apartado rubro, lo que impide realizar un análisis con la totalidad de los datos afectando la calidad de los resultados obtenidos. Es importante remarcar que los datos correspondientes a la mediana del salario son representativos de una población de trabajadores registrados, lo cual no incluiría a los trabajadores informales.

En base a esto, una vez completado el proceso de calidad de datos, se observó que la actividad 'Agricultura, ganadería, caza y servicios relacionados' está asociada a menores salarios que el resto de las actividades, sobre todo en provincias con mayor número de operadores certificados (Figura 5). Esto es consistente con que la agricultura, la ganadería, la caza son industria primaria, mientras que el resto de las actividades se clasifican en la industria secundaria. En términos de cantidad de mano de obra, la industria primaria suele requerir más trabajadores que la industria secundaria (Senado de la Nación, 2011, p. 7), ya que implica actividades que requieren trabajo manual intensivo. Al requerir una mayor cantidad de trabajadores, es razonable pensar que las empresas operadoras pagan salarios menores a sus empleados.

Anexo



Anexo 1. Gráfico de dispersión entre el sueldo promedio y la cantidad de operadores por provincia y por actividad (Clae2). Cada provincia se representa con el color del punto, mientras que la forma de cada dato representa la actividad correspondiente. Los puntos corresponden a 'Agricultura, ganadería, caza y servicios relacionados', las equis a 'Elaboración de productos alimenticios', los cuadrados a 'Elaboración de bebidas', las cruces grandes a 'Elaboración de productos textiles', los rombos a 'Otras industrias manufactureras' y las cruces pequeñas a 'Otros sectores'. El promedio se calculó en base a los salarios de todo el año 2022.



Anexo 2. Gráfico de dispersión entre el sueldo promedio y la cantidad de operadores por departamento y por actividad (Clae2). Los colores de los puntos representan su actividad correspondiente. El promedio se calculó en base a los últimos salarios de 2022.

Referencias bibliográficas

- Centro de Estudios para la Producción, 2022. Salarios promedio y mediano por departamento. Buenos Aires.
- Senado de la Nación, 2011. Proyecto de ley de promoción de los productos orgánicos. Buenos Aires.