Laboratorio de datos
TP-01

Otermín Juana, Quispe Rojas Luis Enrique , Vilcovsky Maia
Facultad de ciencias exactas y naturales

Resumen

Este trabajo analiza la correlación entre los flujos monetarios netos de Inversión Extranjera Directa (IED) y la cantidad de sedes que Argentina tiene en el exterior. Para ello, se utilizó información detallada sobre el IED y las sedes argentinas a nivel global, sometiendo los datos a un proceso de limpieza para asegurar su relevancia y calidad antes de su análisis. Mediante técnicas de visualización de datos se pudo observar que, para el año 2022, un aumento en el número de sedes en un país está asociado con un incremento en la inversión directa recibida por Argentina. Cuando se analizaron los flujos de 2018 a 2022 se observan tres excepciones: Oceanía tiene pocas sedes pero alto flujo monetario; América del Norte, pese a tener menos sedes, lidera en flujos de inversión; y América del Sur, aunque tiene la mayor cantidad de sedes, ocupa el cuarto lugar en flujo monetario.

A. Introducción

La Inversión Extranjera Directa (IED) y las representaciones de un país en el exterior son dos aspectos fundamentales en las relaciones internacionales y la economía global. La IED representa los flujos monetarios resultantes de las inversiones directas de empresas extranjeras en activos físicos o financieros de un país, mientras que las representaciones exteriores, como embajadas y consulados, son vitales para promover los intereses diplomáticos, comerciales y culturales de un país en el extranjero.

En este estudio, nos enfocaremos en analizar la posible relación entre los flujos monetarios netos de IED y la cantidad de sedes que Argentina tiene en el exterior. Este trabajo es relevante dada la importancia estratégica de la IED para la economía argentina y el papel crucial de las representaciones en el extranjero en la promoción de los intereses del país.

Se contó con fuentes de datos detalladas sobre el flujo de IED por país y año, así como información sobre las sedes de Argentina en el exterior, incluyendo ubicación, tipos de secciones y otros detalles relevantes. Además, se dispuso de datos sobre los países involucrados, facilitando la conexión entre los dos conjuntos de datos.

Para lograr el objetivo de este trabajo, se realizó una limpieza de las bases de datos, seleccionando únicamente la información relevante para nuestro análisis. Se emplearon técnicas de visualización de datos para explorar y representar gráficamente la relación entre la IED y la presencia de Argentina en el exterior. A través de este análisis, se buscó determinar si existe una relación significativa entre estos dos factores claves en las relaciones internacionales y económicas de Argentina.

B. Procesamiento de Datos

Se descargaron y observaron las tablas correspondientes a los datos de las Representaciones Argentinas en el exterior (Ver referencia [1]).

Para la tabla denominada LISTA _SEDE se observó que pertenece a primera forma normal ya que posee atributos atómicos. Identificando como clave primaria *sede_id*, se concluyó que se encuentra en segunda forma normal. Esto se debe a que la clave primaria consta de un solo atributo, lo que garantiza que todos los atributos no primos dependan completamente de ella. Y que no se encuentra en tercera forma normal debido a la presencia de dependencias transitivas. Un ejemplo de esto es el atributo *pais_castellano*, el cual no depende directamente de *sede_id*, sino de *pais_iso_2*, que a su vez depende de *sede_id*.

Para la tabla denominada LISTA _SECCIONES se observó que no pertenece a primera forma normal, ya que por ejemplo, el atributo *telefono_principal* posee los siguientes valores: " 56 9 9842 3177 / 56 9 9223 2419 " que no son atómicos. Por lo tanto tampoco se encuentra en segunda y tercera forma normal.

Para la tabla denominada LISTA _SEDES _DATOS se observó que no pertenece a primera forma normal, ya que posee valores no atómicos. Un ejemplo de esto es para el atributo *redes_sociales* que posee el siguiente valor: "https://www.facebook.com/ConsuladoArgent.. // http://www.instagram.com/consulado.argentino.bonn //". Debido a que la tabla no se encuentra en primera forma normal tampoco se encuentra en segunda y tercera forma normal.

Se siguieron algunos procesos para aumentar la calidad de los datos, ya que los dataset contaban con varios de estos problemas. A continuación se muestra un problema por cada fuente.

LISTA _SEDES:

Se afecta el atributo de calidad Relevancia. Esta tabla posee como atributo *estado* que se refiere al estado de una sede, el cual puede tomar los valores *activo* o *inactivo*. Se consideró que este atributo no es relevante para nuestro objetivo ya que si una sede está inactiva no nos importan sus datos.

Este es un problema asociado al modelo, ya que el diseño del modelo de datos incluye el atributo *estado* cuando no es necesario. Y no es un problema de instancia ya que los datos son actuales, consistentes y claros.

Para tener una medida concreta acerca de la magnitud del problema se usó el método de GQM: Queremos analizar la relevancia del dato *estado* asociado a la tabla sedes.

Goal: El dato correspondiente al Estado de cada sede debe ser *Activo* para poder garantizar que la sede está operando físicamente en este momento.

Question: ¿Que porcentaje de sedes inactivas hay en mi Dataframe?

Metric: Cantidad total de datos de mi columna *estado* - cantidad de datos correspondientes a un estado "activo" y luego pasarlo a porcentaje.

Calculo: 164 - 162 = 2 sedes "inactivas"

Porcentaje de sedes "inactivas": 2/164 * 100 = 1,22 %

SECCIONES _DATOS:

Se afecta el atributo de calidad Consistencia. Ya que hay contradicciones entre los datos almacenados. Para un mismo atributo hay un valor que es una URL y otro valor que es un usuario que comienza con un "@".

Este es un problema asociado a instancia. Probablemente estos datos provienen de distintas fuentes y debido a eso no tienen consistencia.

Para tener una medida concreta acerca de la magnitud del problema se usó el método de GQM: Queremos analizar la consistencia de los datos de redes sociales en la tabla de secciones.

Goal: Asegurar que los datos de redes sociales en la tabla de secciones sean consistentes en términos de formato y estructura.

Question:¿Qué porcentaje de datos no tienen la estructura de URL?

Metric: Total de filas en *redes_sociales* - filas que contengan .com (formato URLs) y luego pasarlo a porcentaje.

Calculo = 266 - 230 = 36 datos con distinta estructura

Porcentaje de datos sin el formato URL: 36/266 * 100 = 13,53 %

SECCIONES:

Se afecta el atributo de calidad Completitud, hay datos en la columna telefono_principal que contienen nulls.

Este problema corresponde a instancia. El modelo de datos es el correcto, el problema es que no todos los datos tienen un valor.

Para tener una medida concreta acerca de la magnitud del problema se usó el método de GQM: Queremos analizar la completitud de los datos del atributo *telefono principal* de la tabla secciones.

Goal: Chequear que los datos de $telefono_principal$ en la tabla de secciones sean datos no nulls, para poder sacar información de ellos.

Question: ¿Qué porcentaje de datos null hay en la columna *telefono principal*?

Metric: Porcentaje de datos no nulls en *telefono principal*.

Calculo: 516 - 373 = 143

Porcentaje de datos null: 143/ 516 * 100 = 27,71 %

Se generó un Diagrama Entidad-Relación (DER) que permite modelar los problemas planteados en el presente trabajo. (Figura 1)

En el Diagrama se muestran las entidades fuertes, las entidades débiles, las claves primarias, los atributos y las relaciones entre las entidades. Se consideró como entidad débil a aquellas que no pueden ser identificadas de manera única por sus atributos propios y necesitan hacer referencia a otra entidad (conocidas como entidades propietarias). En nuestro caso tenemos a FLUJOS_MONETARIOS y SECCIONES como entidades débiles con sus correspondientes entidades propietarias (ver 1). Las claves se representan subrayadas.

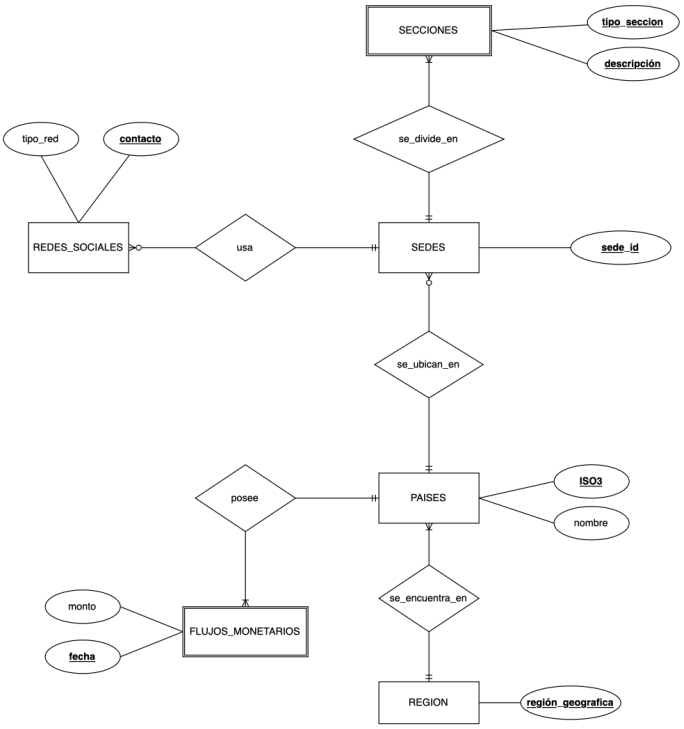

Figura 1: Diagrama Entidad Relación. Un rectángulo representa las entidades fuertes, doble rectángulo representa las entidades débiles, los rombos indican cómo se relacionan las entidades, los óvalos representan los atributos de cada entidad.

Se realizó el mapeo del DER al modelo relacional correspondiente. El mismo se encuentra en tercera forma normal (3FN), es decir, no hay dependencias transitivas entre las clave y los atributos (Figura 2).

En el diagrama del modelo relacional se muestran las entidades con sus dependencias funcionales (en 3FN). Se tomó como claves primarias las superclaves que si se les elimina algún atributo dejan de serlo, es decir, no se pueden sacar atributos de esta sin perder la propiedad de identificación única, además no poseen ningún NULL. Se identificaron en el diagrama como PK. Luego tenemos las Foreign Keys representadas en el diagrama como FK, las mismas hacen referencia a la clave primaria de otra tabla. Mirando el DER 1 vemos que las entidades débiles (flujos _monetarios y secciones) poseen la FK de su entidad propietaria. A SEDE se le incluye la PK de PAISES como FK, esto es debido a la relación uno a muchos (1:N) que poseen estas entidades. Lo mismo pasa con SEDES y REDES _SOCIALES, a esta última se le agrega la PK de SEDES como FK. Y tambien pasa con PAISES y REGION, donde se le agrega la PK de REGION como FK de PAISES.

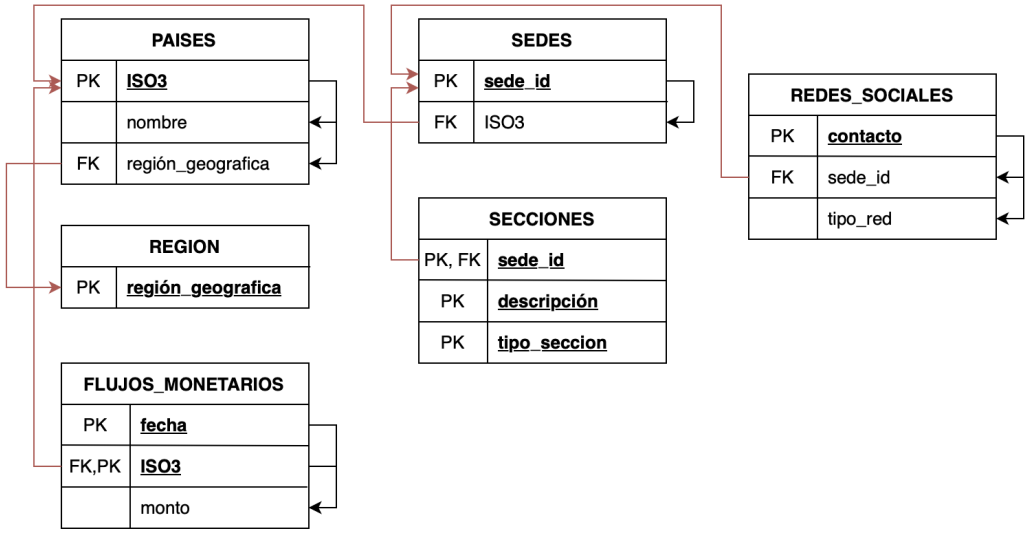

Figura 2: Mapeo del DER al Modelo Relacional. Se representa cada entidad con sus correspondientes claves primarias (PK), Foreign Keys (FK) y atributos. Se representan las dependencias funcionales con las flechas dentro de la misma entidad.

Los tablas originales se fueron limpiando hasta llegar a la estructura planteada en el DER, es por eso que no fue necesario importar los datos a los esquemas vacios.

C. Decisiones tomadas

Se tomaron numerosas decisiones a lo largo de este trabajo, muchas de ellas relacionadas con la calidad de los datos. A continuación, se detallan las mismas para cada DataFrame utilizado.

PAISES: Se seleccionaron los atributos “nombre” y “iso3”, ya que estos no contenían valores NULL y alcanzaban para lograr nuestro objetivo. Luego, se convirtieron los nombres de países a mayúsculas y se reemplazaron caracteres especiales y acentos para asegurar la uniformidad.

LISTA_SECCIONES: Se renombró a SECCIONES. Se conservaron los atributos “sede_id”, “tipo_seccion” y “sede_desc_castellano”, este último se renombró como “descripcion”. La elección de estos atributos se basó en que los demás eran irrelevantes para nuestro objetivo y además presentaban problemas de calidad, como la vigencia de algunos de esos datos.

LISTA_SEDES_DATOS: Se dividió en dos nuevos DataFrames: uno llamado REGION, que incluye el atributo “region_geografica”, y otro llamado REDES_SOCIALES, que incluye los atributos “contacto”, “sede_id” y “tipo_red”. En REDES_SOCIALES, se eliminaron datos que en el atributo “contacto” tenían un valor distinto a todos los demás, es decir, aparecian una sola vez.

FLUJOS: Se renombró a FLUJOS MONETARIOS. Debido a que los datos no se encontraban en un formato accesible, se transpuso el DataFrame y se eliminó la primera fila. Se renombró el atributo “indice_tiempo” a “nombre”, se reemplazaron los valores NULL por

0, y se convirtieron los nombres de países a mayúsculas. Además, se reemplazaron ciertos nombres de países por sus equivalentes en el DataFrame PAISES.

LISTA_SEDES: Se renombró a SEDES. Se eliminaron las filas donde el atributo "estado" era "Inactivo", ya que las sedes inactivas no eran relevantes para nuestro objetivo. Finalmente, se conservaron los atributos "sede_id" y "pais_iso_3", este último renombrado a "iso3".

D. Análisis de datos

Como primera aproximación al problema se dispuso crear una tabla que resumiera la cantidad de sedes, cantidad de secciones y flujo promedio por país en el último año del que se tenía registro (2022). Las primera filas se observan en la figura 3.

	pais	sedes	secciones promedio	IED 2022 (M U$S)
0	BRASIL	11	1.09	86050.358974
1	ESTADOS UNIDOS DE AMERICA	9	1.33	285057.000000
2	URUGUAY	8	1.00	3838.745569
3	BOLIVIA	7	1.14	-26.358420
4	CHILE	7	1.14	19786.004980
5	ESPAÑA	7	1.14	34811.072593
6	CHINA	5	1.60	189132.410000
7	ALEMANIA	4	1.25	11053.395870
8	CANADA	4	1.25	52633.196446
9	ITALIA	4	1.50	19946.842119
10	PARAGUAY	4	1.00	473.529699
11	MEXICO	3	1.33	35291.618356

Figura 3: Tabla para cantidad de sedes, promedio de secciones y flujos promedio por pais para 2022. Los bigotes aparecen truncados debido a la presencia de valores negativos imposibles de visualizar usando la escala logarítmica.

A primera vista la cantidad de sedes no tiene relación con el promedio de secciones, esto se reafirma si vemos la dispersión de esta columna '0.486084'. Los valores de los promedios están muy juntos entre sí y el máximo sería 3.

La relación entre flujo y cantidad de sedes es un poco más difícil de visualizar. Para esto se realizó el gráfico que se muestra en la figura 4.

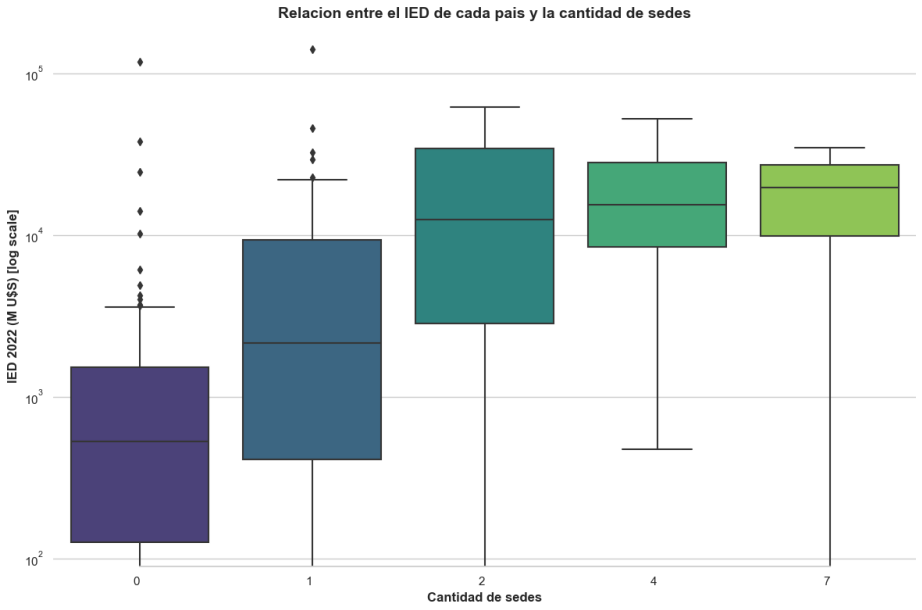

Figura 4: Gráfico que muestra el flujo monetario para 2022 de cada país vs la cantidad de sedes.

Se graficó la cantidad de sedes por cada país en función del flujo monetario para el año 2022. Fue necesario el uso de una escala logarítmica para el flujo debido a la diferencia de escalas que se tenía. Además, para los valores de *sedes* que tenían solo un país se optó por no graficarlos. Lo que se observa del gráfico 4 es que al parecer parece existir una relación creciente entre estas dos magnitudes para la mediana. Lo que indicaría que a mayor cantidad de sedes, existiría una mayor cantidad de flujos de inversión directa.

Mediante otro enfoque podríamos visualizar estos mismos datos por región geográfica. Para esto se tabuló esta magnitud en forma descendente en la tabla 5 (los datos corresponden al año 2022).

Index	Región geográfica	Países Con Sedes Argentinas	Promedio IED 2022 (M U$S)
0	AMÉRICA DEL NORTE	3	124327
1	OCEANÍA	2	34584.1
2	ASIA	23	25863.4
3	AMÉRICA DEL SUR	11	14150
4	EUROPA OCCIDENTAL	18	7693.2
5	EUROPA CENTRAL Y ORIENTAL	8	6162.08
6	ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	2868.69
7	AMÉRICA CENTRAL Y CARIBE	13	1464.98
8	ÁFRICA SUBSAHARIANA	7	1340.52

Figura 5: Tabla que representa la cantidad de países en que Argentina tiene al menos una sede y el promedio del IED 2022 de esos países, agrupados por región geográfica.

Esta consulta parece indicar una discrepancia con lo antes mencionado para los distintos países. Pero cabe resaltar, que esta consulta aglomera distintos países por región. Por este

motivo se decidió ampliar la ventana temporal y graficar el número de sedes en un gráfico de barras. Esto se muestra en la figura 6.

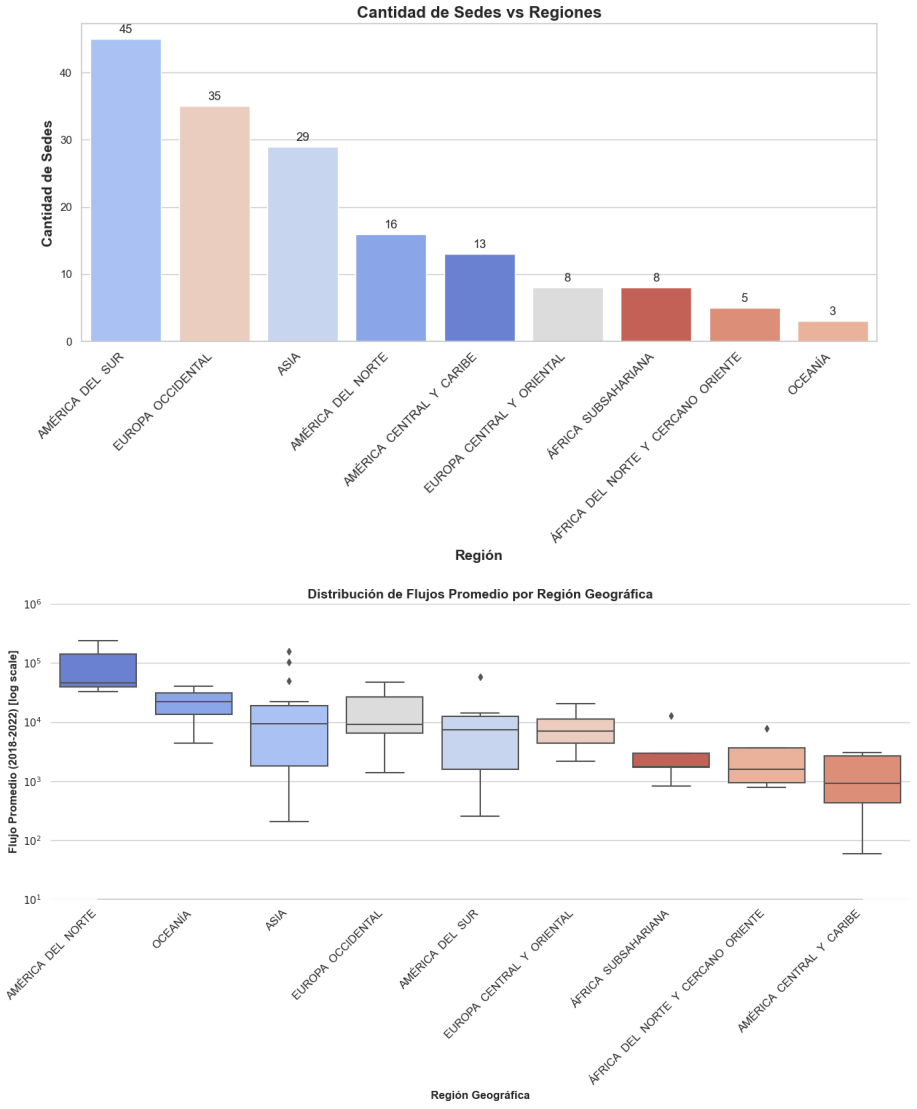

Figura 6: El primer gráfico muestra la cantidad de sedes vs las regiones. El segundo gráfico muestra la distribución del flujo promedio por región.

Se identifica 3 casos excepcionales que no corresponderían con la afirmación que a mayor cantidad de sedes mayor cantidad de flujo. Oceanía era el último en cantidad de sedes pero el segundo monetario. América del Norte pasa de ser el cuarto en cantidad de sedes a ser el primero en flujos de inversión. Y por último América del Sur, el de mayor cantidad de sedes, sería el cuarto en flujos monetarios.

Para las demás regiones geográficas la discrepancia entre cantidad de sedes y flujos monetarios parecía seguir la tendencia. A mayor cantidad de sedes argentina el flujo de inversión directa también lo sería.

E. Conclusion

En conclusión, el presente tuvo como objetivo establecer una relación entre los flujos monetarios netos de Inversión Extranjera Directa (IED) y la cantidad de sedes argentinas en distintos países de una base de datos gubernamental. Se concluyó, para el año 2022, que a medida que aumenta el número de sedes de Argentina en un país, también tiende a incrementarse el monto de IED. Pero cuando se realizó el análisis para 2018 a 2022 hubo tres excepciones: Oceanía tiene pocas sedes pero alto flujo monetario; América del Norte, pese a tener menos sedes, lidera en flujos de inversión; y América del Sur, aunque tiene la mayor cantidad de sedes, ocupa el cuarto lugar en flujo monetario.

Referencias

[1] https://datos.gob.ar/dataset/exterior-representaciones-argentinas

[2] https://datos.gob.ar/dataset/sspm-flujos-monetarios-netos-inversion-extranjera-directa/archivo/sspm_337.1

[3] https://gist.github.com/brenes/1095110