Trabajo Práctico – 01: Relación entre los flujos monetarios netos de IED de cada país y cantidad de sedes en el exterior que tiene Argentina.

Dramis Agustín; Rostan Marcos; Rozenblit Valentín

Laboratorio de datos – 1er. Cuatrimestre 2024

Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

Resumen

En este trabajo práctico se intentó encontrar la relación entre el IED de cada país y la cantidad de sedes en el exterior que tiene Argentina en dicho país. Para ello, se realizó (en los archivos .csv dados por los profesores) un modelo de datos (DER), y se mapeó en el modelo relacional, verificando que las tablas estén en tercera forma normal y mejorando la calidad de datos de las tablas originales. Posteriormente, se realizó una análisis exploratorio de datos en el que se concluyó que a medida que aumentan las sedes argentinas en un país, aumenta la inversión de dicho país hacia Argentina.

Introducción

El objetivo fue ver si existe alguna relación entre los flujos monetarios netos de Inversión Extranjera Directa (IED) de cada país y la cantidad de sedes en el exterior que tiene Argentina en dicho país.

Para ello se realizó un DER para representar visualmente las entidades y las relaciones entre ellas. Luego, se mapeó el DER al modelo relacional para organizar la información en tablas, donde cada fila representa una instancia de la entidad y cada columna un atributo. Previamente, a esas tablas generadas, se verificó que todas estén en tercera forma normal para reducir la redundancia y la dependencia de datos.

Para garantizar la efectividad y confiabilidad del proceso, se realizó una mejora de calidad de datos, utilizando la técnica GQM que consiste en un objetivo claro, preguntas para evaluar si se están logrando esos objetivos y las métricas que se utilizan para responder a esas preguntas de manera efectiva.

Una vez ya realizado todo lo anterior, se hicieron consultas SQL para la generación de las nuevas tablas del modelo relacional y posteriormente para el análisis de datos. A esas consultas se la presentaron como gráficos para poder tener una visualización de aquellos datos y así tomar conclusiones acerca de la relación entre el flujo IED de cada país y la cantidad de sedes en el exterior que tiene Argentina.

Procesamiento de datos

Las 3 tablas de “Representaciones Argentinas” se encuentran de la siguiente forma normal:

- Lista-sedes-datos.csv: No se encuentra en ninguna forma, ya que NO es FN1 porque tiene, por ejemplo, en la columna “circunscripción” valores con coma, es decir, viola el primer principio de que cada celda debe contener un único valor atómico.

- Lista-sedes.csv: Está en FN1 porque los valores de las columnas son atómicos. Está en FN2 porque todos los atributos NO clave dependen de la clave completa {sede_id, sede_desc_castellano}. Y está en FN3 ya que no hay dependencia transitiva en la estructura de la tabla.

• Lista-secciones.csv: No se encuentra en ninguna forma, NO es FN1 porque tiene en la columna "sede_desc_castellano" valores NO atómicos.

Para aumentar la calidad de datos se utilizó el método GQM (Goal, Question, Metric), que consiste en fijar un objetivo claro, hacer preguntas específicas y responder esas preguntas con métricas. De la siguiente forma:

► Lista-sedes-datos.csv:

- El atributo de calidad afectado es **consistencia**. Es un problema de instancia, porque no todas las instancias de los datos se encuentran sincronizadas.
- Objetivo: El dato correspondiente al código postal sea consistente.
- Pregunta: ¿Cuántos datos correspondientes al nombre del titular aparecen vacíos, con ceros, con guiones o con descripción con letras?
- Métricas:
  i. M1: Cantidad de datos en código_postal que sean ceros.
  ii. M2: Cantidad de datos en código_postal que estén vacíos
  iii. M3: Cantidad de datos en código_postal que sean guiones.
  iv. M4: Cantidad de datos en código_postal que sean letras.
  v. M5: Suma de la cantidad total de datos en código_postal que no sean números.
- Criterio utilizado para corregir los datos: Implementar una regla que cuando no se pueda obtener el código postal, el dato correspondiente será "Sin identificar".

Al implementar esta regla, el resultado de las métricas se vuelve 0. Quiere decir que esa columna será consistente.

► Flujos-monetarios-netos-inversion-extranjera-directa.csv:

- El atributo afectado es **Disponibilidad**. Es un problema de modelo porque está mal organizada la base de datos.
- Objetivo: Los datos estén accesibles.
- Pregunta: ¿Se puede acceder fácilmente a los flujos por fecha?
- Métricas:
  i. M1: Cantidad de clics necesarios del usuario para encontrar los datos de un país específico.
  ii. M2: Promedio de tiempo de búsqueda de los datos de un país específico.
- Criterio: Trasponer la tabla, para que las fechas queden como encabezado. La cantidad de clics será menor y el promedio de tiempo de búsqueda bajará, esto quiere decir que habrá una mejor accesibilidad de los datos.

► Lista-secciones.csv:

- El atributo afectado es **Completitud** y es un problema de instancia, ya que es un problema en una etapa del proceso.
- Objetivo: Los datos correspondientes al teléfono principal estén completos.
- Pregunta: ¿Cuántos datos correspondientes al número del teléfono principal están vacíos?
- Métricas: M1: Cantidad de datos en teléfono_principal que estén vacíos.
- Criterio: Implementar una regla en donde en vez de dejar vacío el dato, sé pondrá "Sin asignar". Esto genera que la cantidad de datos en teléfono_principal sea 0, entonces se garantiza que el campo teléfono_principal esté completo.

Se generó un Diagrama Entidad-Relación contemplando las variables relevantes al problema (Figura 1). La cantidad de secciones de una sede diplomática se tomó simplemente como un atributo, en lugar de una entidad propia, al no tener como objetivo ninguna propiedad de ellas. A partir del diagrama se generó un modelo relacional (Figura 2). Al ser todas las relaciones "uno a muchos", no se incluyeron tablas para las relaciones entre entidades, incluyendo en su lugar foreign keys en todas las tablas a excepción de país.

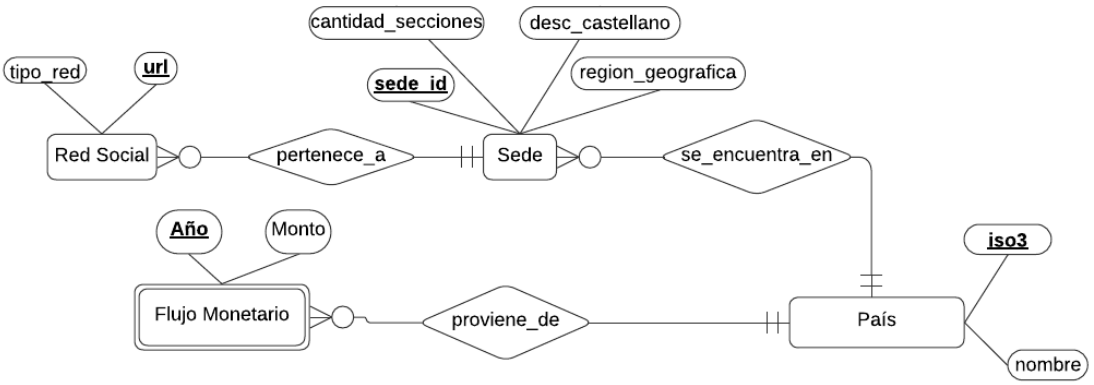

Figura 1 - DER utilizado en el trabajo práctico

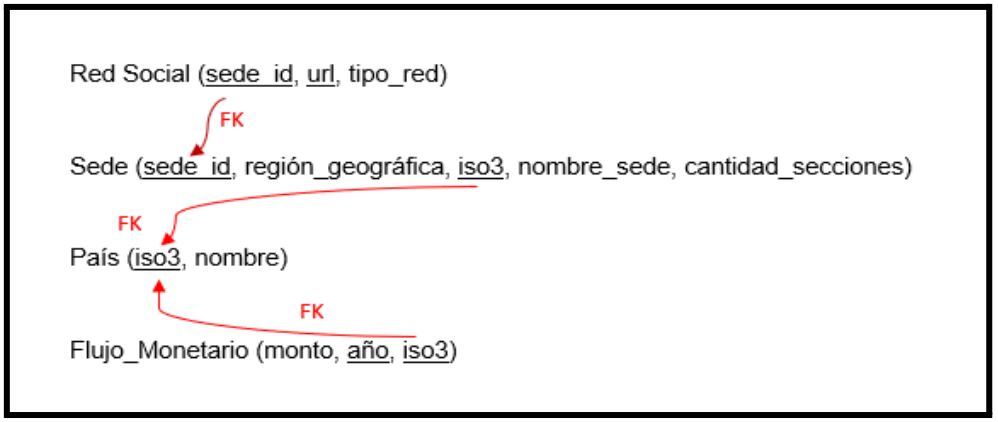

Figura 2 - Mapeo al modelo relacional

Decisiones tomadas

- Dado que en *países.csv* la columna *iso2* presentaba un dato faltante (correspondiente a Namibia) mientras que *iso3* se encontraba completa, se utilizó esta última como identificador de países.

- Al estar los nombres de los países escritos de una forma “simplificada” en la tabla *flujos-monetarios-netos-inversion-extranjera-directa.csv*, se generó una nueva columna en la tabla *paises.csv* con esta forma.
- En cuanto a las redes sociales, se decidió considerar como válidas únicamente las que incluyeran la secuencia “.com”, y descartar aquellas con forma “@%”, al no poder asignarles un tipo de red social específico.
- Para extraer la red social del URL, se realizó primero un análisis exploratorio, observando las distintas secuencias de caracteres previas a “.com”, determinando la existencia de 6 redes sociales distintas: Instagram, Facebook, Twitter, Youtube, Linkedin y Flickr. El tipo de red se extrajo entonces mediante el uso de sentencias CASE WHEN de sql, considerando la presencia de secuencias del estilo “%[nombre_red]%” en minúsculas. Una limitación de esta metodología es que, de incorporarse datos, se debería realizar de nuevo la exploración inicial para contemplar la incorporación de redes distintas a las observadas.

Análisis de datos

Los países presentaron entre 1 y 5 sedes diplomáticas, con entre 1 y 14 secciones en promedio. La inversión contuvo valores tanto positivos como negativos (implicando una desinversión), yendo desde -67.340 M U$D hasta 189.132 M U$D (anexo como archivo .csv, head en Tabla 1). Al graficar la relación entre la IED y la cantidad de sedes se observó una tendencia a mayores valores niveles de inversión provenientes de países con mayor cantidad de sedes diplomáticas. Sin embargo, debe tomarse en cuenta que a excepción de Brasil (con 5 sedes), China y España (con 3 sedes cada uno), todos los países presentaron entre una y dos sedes, por lo que esta tendencia podría modificarse ampliamente si se incorporan más países con mayor cantidad de sedes (por ejemplo en un año futuro).

index	pais	sedes	secciones_promedio	IED_2022_M_U$S
0	Brasil	5	3.6	86050.4
1	China	3	5.66667	189132
2	España	3	5.33333	34811.1
3	Alemania	2	7	11053.4
4	Australia	2	4	61629.5
5	Bolivia	2	8	-26.3584
6	Bélgica	2	5.5	-1710.15
7	Canadá	2	3.5	52633.2
8	Ecuador	2	5	788.058
9	India	2	3.5	49354.6
10	Israel	2	2.5	27760.1

Tabla 1 - Sedes, secciones promedio e IED durante el año 2022 para cada país (head).

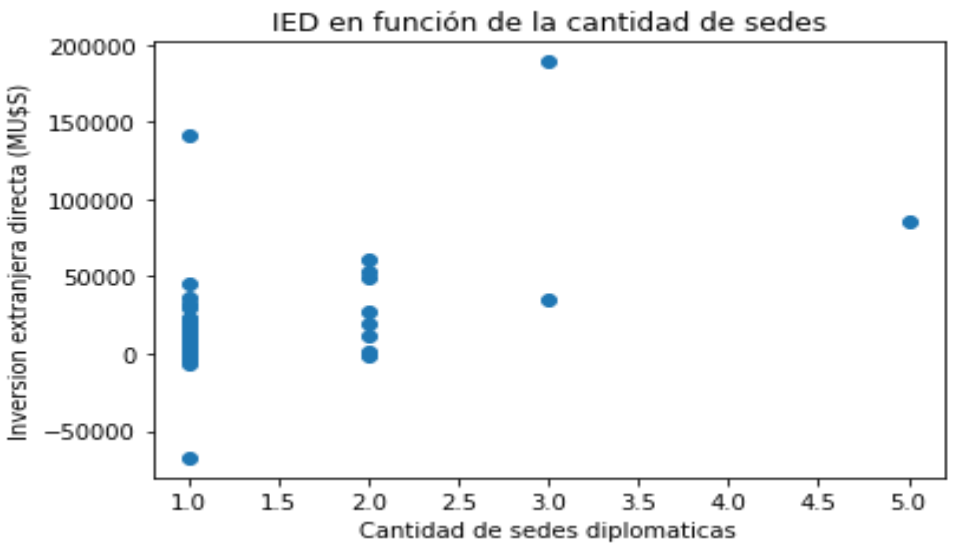

Figura 3 -Relación entre IED y cantidad de sedes diplomáticas para cada país

Al agrupar por región geográfica, se observó que América del Norte, Asia y Oceanía fueron las regiones de las que provino mayor IED durante 2022 (Tabla 2). Esta tendencia general se mantuvo al extender el período analizado a 2018-2022 (Figura 4, y detalle sin outliers en Figura 5). Cabe resaltar que si bien América del Norte cuenta con 3 países con sedes argentinas, la IED de ese período provino exclusivamente de Canadá. La región con más sedes diplomáticas argentinas resultó ser Asia (Figura 6), pero a escala regional no se evidenció una relación entre cantidad de sedes e inversión.

Index	region_geografica	Paises_con_sedes_Argentinas	prom_IED_2022
0	AMÉRICA DEL NORTE	3	52633.2
1	ASIA	20	51646.2
2	OCEANÍA	2	43599.2
3	AMÉRICA DEL SUR	9	34646
4	EUROPA CENTRAL Y ORIENTAL	7	11123.9
5	EUROPA OCCIDENTAL	18	11100.2
6	ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	4751.55
7	AMÉRICA CENTRAL Y CARIBE	12	1366.44
8	ÁFRICA SUBSAHARIANA	6	1091.39

Tabla 2 - Cantidad de países con sedes diplomáticas argentinas por región geográfica e IED promedio de esos países en 2022

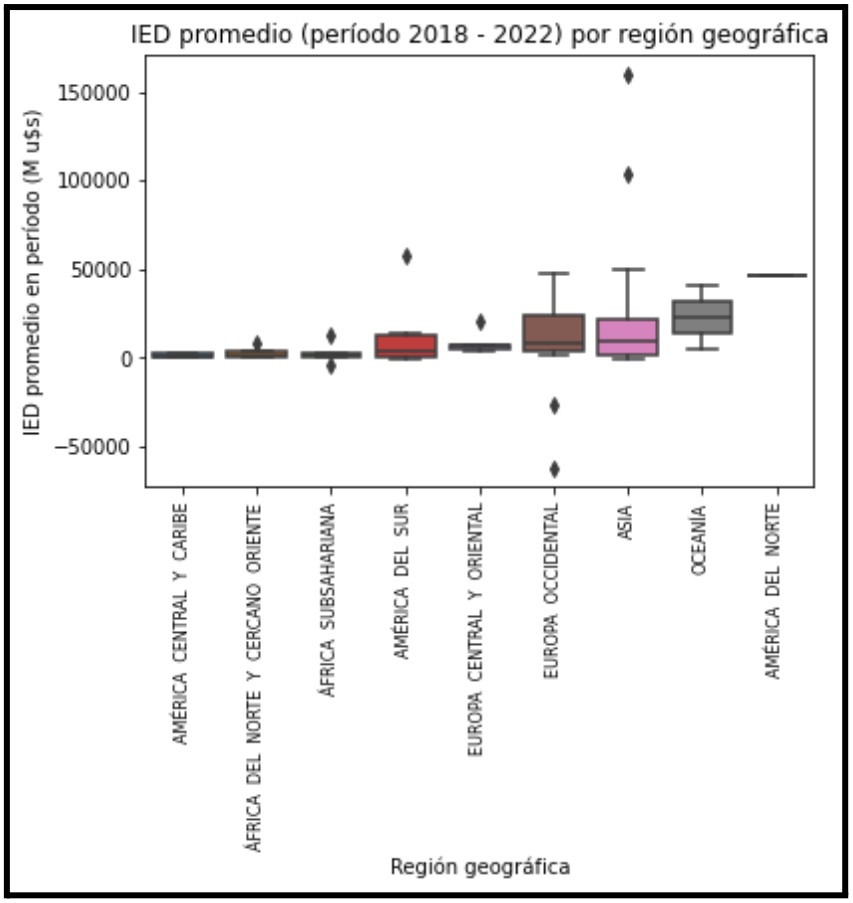

Figura 4 - IED promedio para el período 2018-2022 por región geográfica.

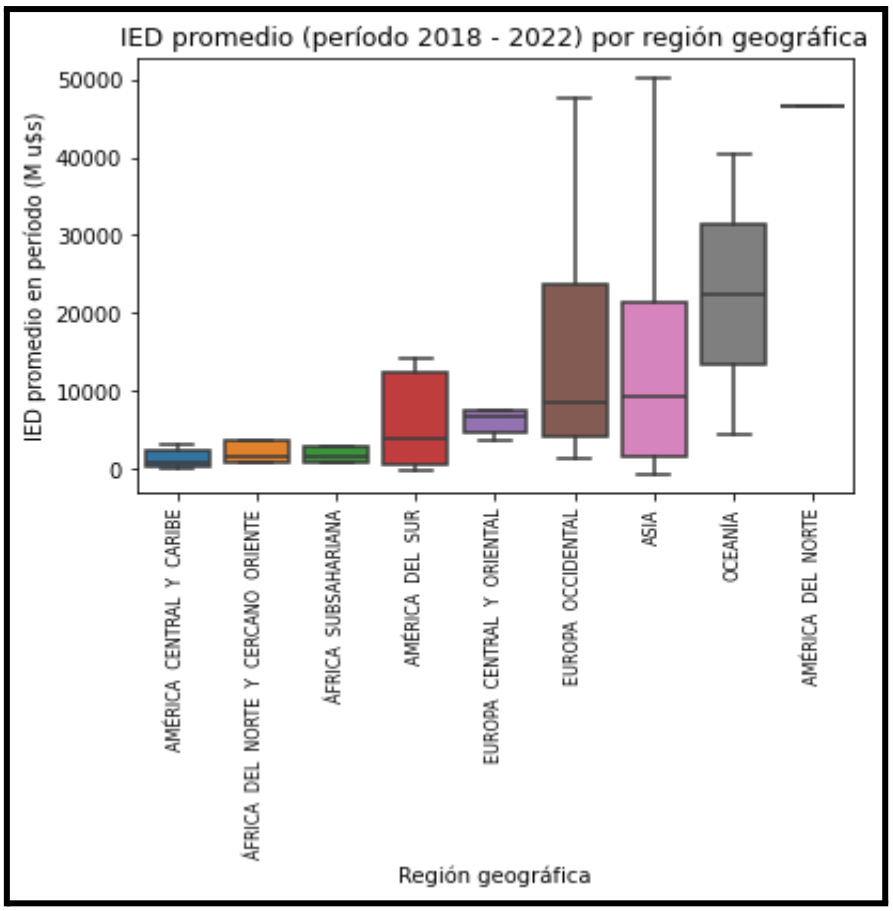

Figura 5 - IED promedio para el período 2018-2022 por región geográfica, sin mostrar los outliers

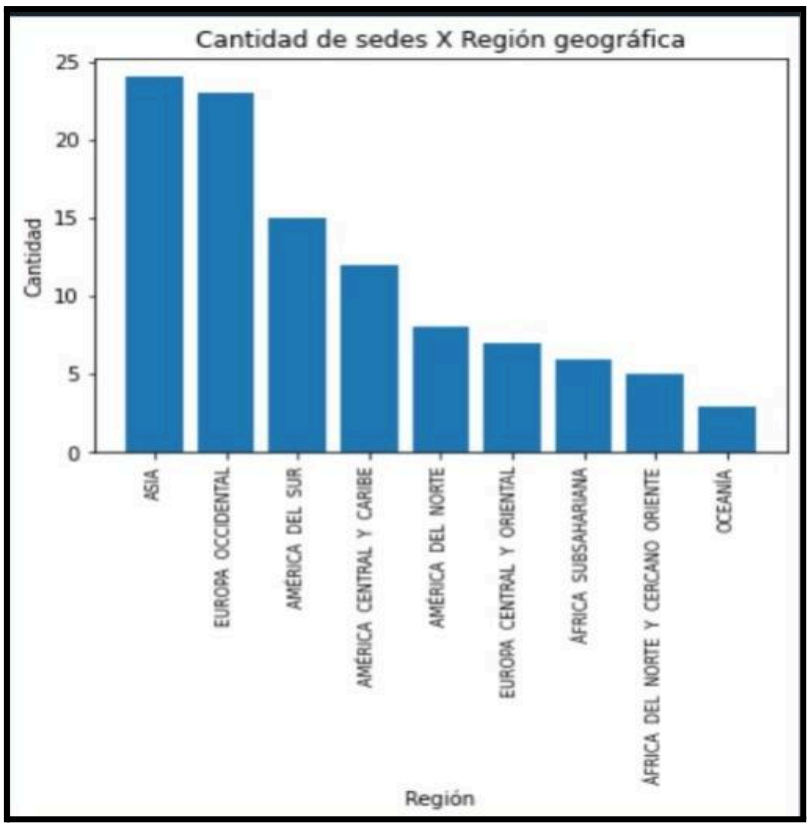

Figura 6 - Cantidad de sedes diplomáticas argentinas por región geográfica.

Por otro lado, se analizaron los medios de comunicación que utilizan las distintas sedes diplomáticas. Para ello se consideraron las redes sociales de sedes diplomáticas (archivo .csv anexo, head en tabla 3). Las sedes diplomáticas presentaron entre 0 (en el caso de Serbia) y 6 (en Bélgica y Estados Unidos) redes sociales distintas (archivo .csv anexo, head en tabla 4).

index	nombre	nombre_sede	tipo_red	url
0	Alemania	Embajada en Alemania	Facebook	https://www.facebook.com/ArgEnAlemania/
1	Algeria	Embajada en Argelia	Facebook	facebook.com/ArgentinaEnArgelia
2	Algeria	Embajada en Argelia	Instagram	https://instagram.com/argenargelia
3	Algeria	Embajada en Argelia	Twitter	https://twitter.com/ARGenArgelia
4	Angola	Embajada en Angola	Facebook	https://www.facebook.com/ArgentinaEnAngola/
5	Angola	Embajada en Angola	Instagram	https://www.instagram.com/embargentinaenangola/
6	Armenia	Embajada en Armenia	Facebook	https://www.facebook.com/ArgentinaEnArmenia
7	Armenia	Embajada en Armenia	Instagram	https://www.instagram.com/arginarmenia/
8	Armenia	Embajada en Armenia	Twitter	https://twitter.com/ARGinArmenia
9	Australia	Consulado General en Sidney	Facebook	https://www.facebook.com/ArgentinaEnSidney/

Tabla 3 - Redes sociales de sedes diplomáticas argentinas (head)

Index	pais	cantidad_red
0	India	3
1	Bélgica	6
2	Ciudad del Vaticano	4
3	Malasia	1
4	Alemania	1
5	Pakistán	1
6	Trinidad y Tobago	1
7	Uruguay	3
8	Suecia	3
9	Filipinas	3
10	Barbados	1
11	Francia	1
12	Dinamarca	1
13	Líbano	1

Tabla 4 - Cantidad de redes sociales distintas por país (head)

Conclusiones

En este trabajo práctico se analizó la relación entre la Inversión Extranjera Directa proveniente de un país y la cantidad de sedes diplomáticas argentinas en dicho país. Se pudo observar una tendencia en este sentido, con mayores medianas de inversiones provenientes de países con mayores cantidades de sedes. Sin embargo, la escasa variedad de cantidades de sedes debe ser tomada en cuenta. A futuro, podrían considerarse distintas variables de las sedes diplomáticas, como la cantidad de secciones o los tipos de redes sociales en su asociación con la inversión.