

Práctica 3: Explotación de datos

Patricia Lázaro Tello

Índice

1. Introducción	2
2. Creación del modelo OLAP	2
2.1. Creación del proyecto e importación de la base de datos	2
2.2. Origen de datos	3
2.3. Vista del origen de datos	6
2.4. Creación e implementación de los cubos	8
2.5. Configuración de dimensiones conformadas	14
2.6. Jerarquías y dimensiones	17
2.7. Procesado y resolución de errores	20
3. Explotación del modelo OLAP	24
3.1. Biodiversidad en Cantabria	24
3.2. Producción en E.E.U.U.	26
3.3. <i>Top 5</i> áreas protegidas	29
3.4. Ciudades y Comunidades sostenibles	35
3.5. Gestión de aguas residuales en la C. Valenciana	37
3.6. Comparativa del área protegida	40

1. Introducción

Hasta ahora, se han definido los requerimientos del *data warehouse*, se han analizado las fuentes de datos y se ha diseñado el modelo conceptual, lógico y físico del almacén de datos. También se ha llevado a cabo la creación de las tablas diseñadas y la carga de datos en el modelo multidimensional.

El último apartado a tratar es la **explotación de los datos**: para poder obtener conocimiento de los mismos, es necesario crear un modelo OLAP, que permitirá realizar consultas de forma ordenada, estandarizada y rápida.

2. Creación del modelo OLAP

A continuación se procede a detallar el proceso de creación y configuración del modelo OLAP, que incluye desde los pasos básicos como crear el proyecto y configurar el origen de datos, hasta la conformación de dimensiones y creación de jerarquías dentro de las mismas.

2.1. Creación del proyecto e importación de la base de datos

En primer lugar, se ha de crear el proyecto y configurar la base de datos a utilizar, que corresponde con «DEST_plazarotello».

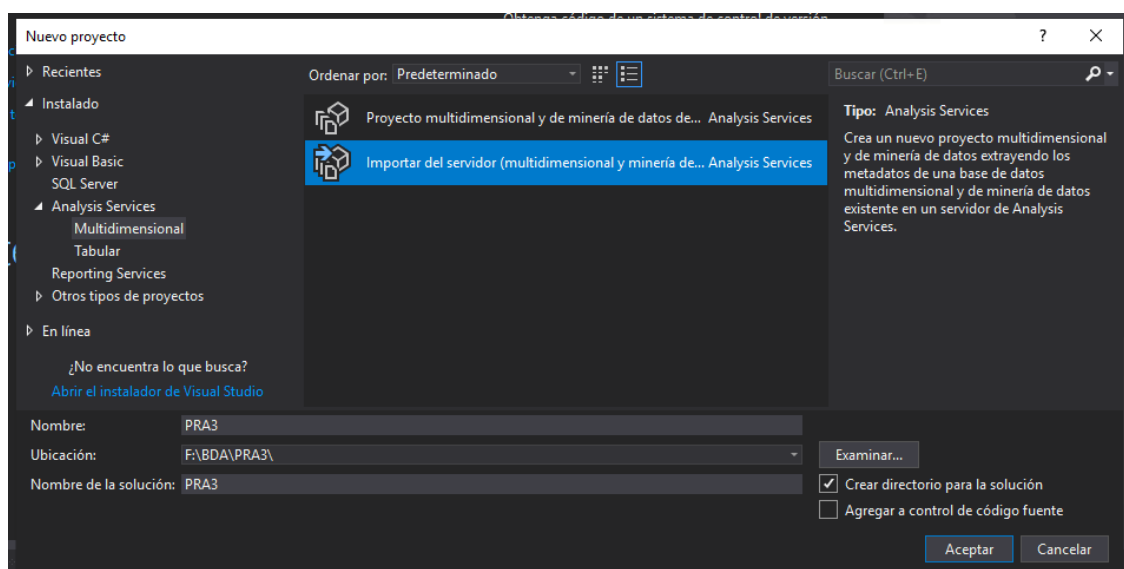
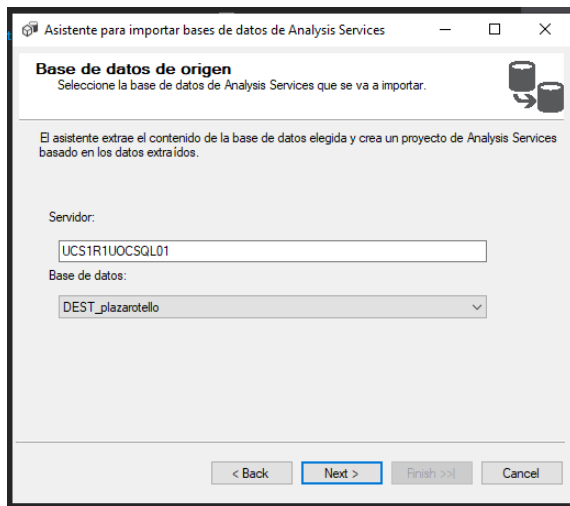
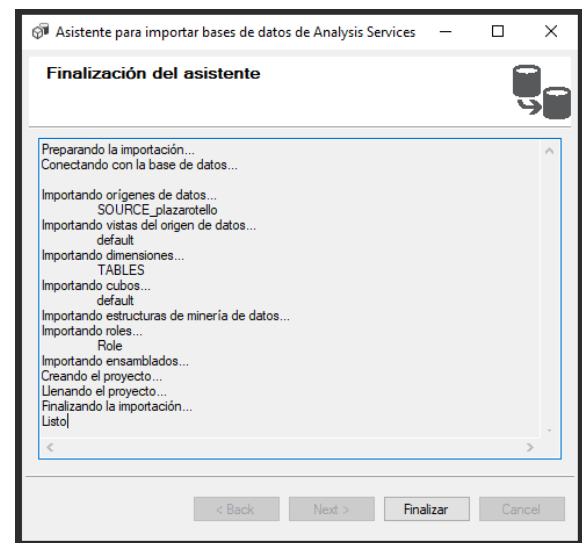


Figura 1: Creación de un proyecto *Analysis Services*



(a) Configuración de la base de datos de origen



(b) Proyecto creado con éxito

2.2. Origen de datos

A continuación se configura el servidor de implementación:

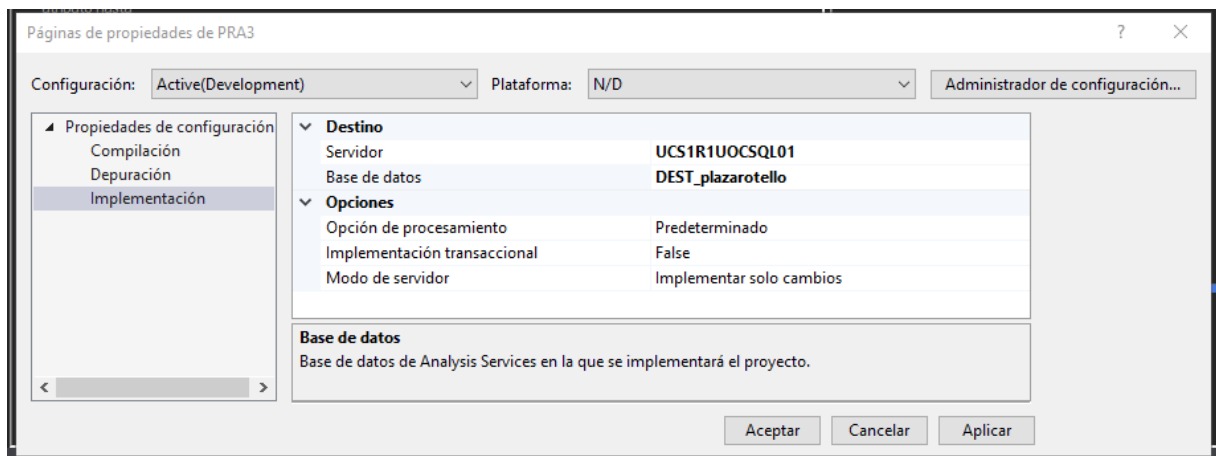


Figura 2: Servidor y base de datos de destino

Así como la conexión a la base de datos y el mecanismo de impersonalización:

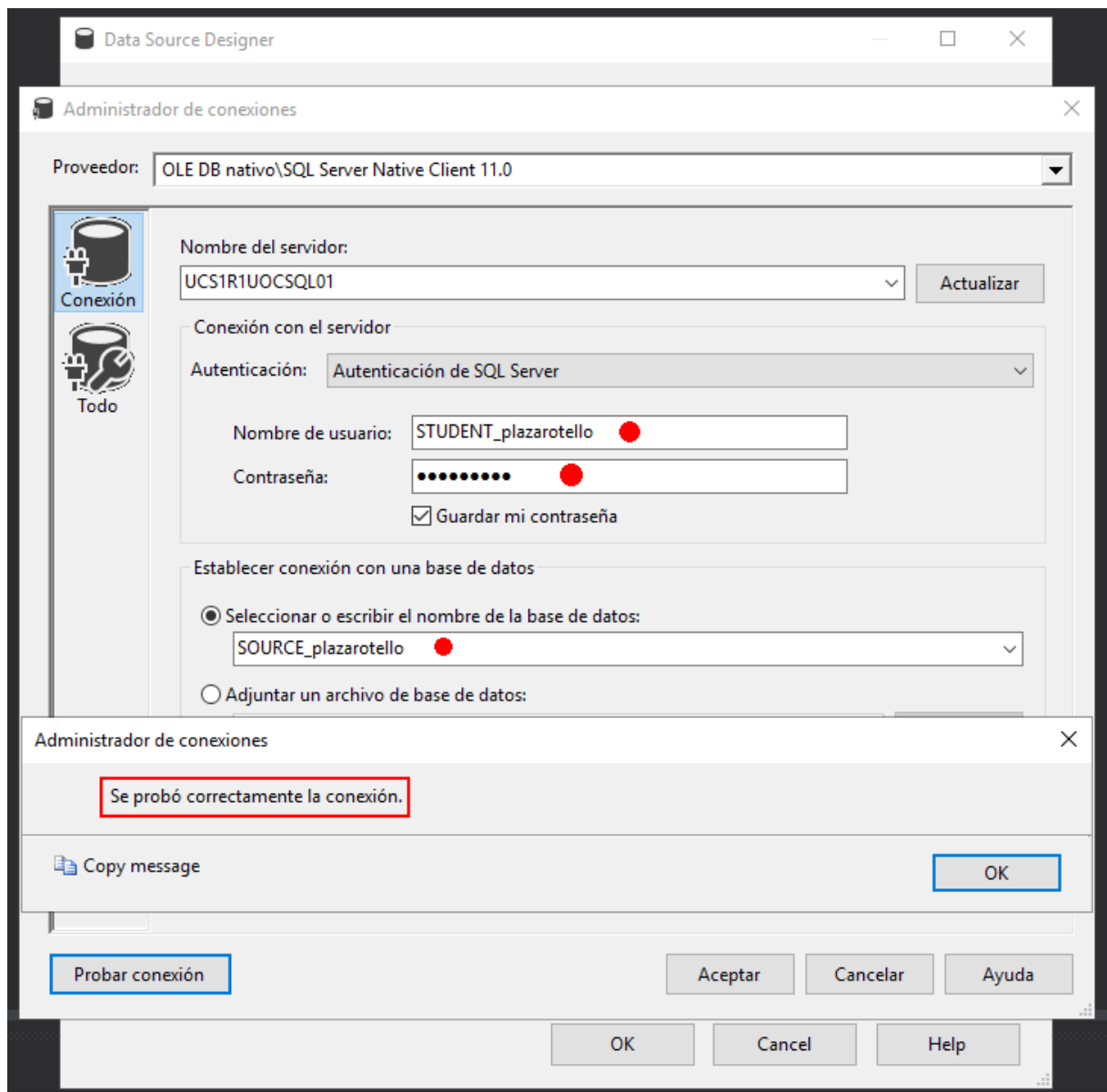
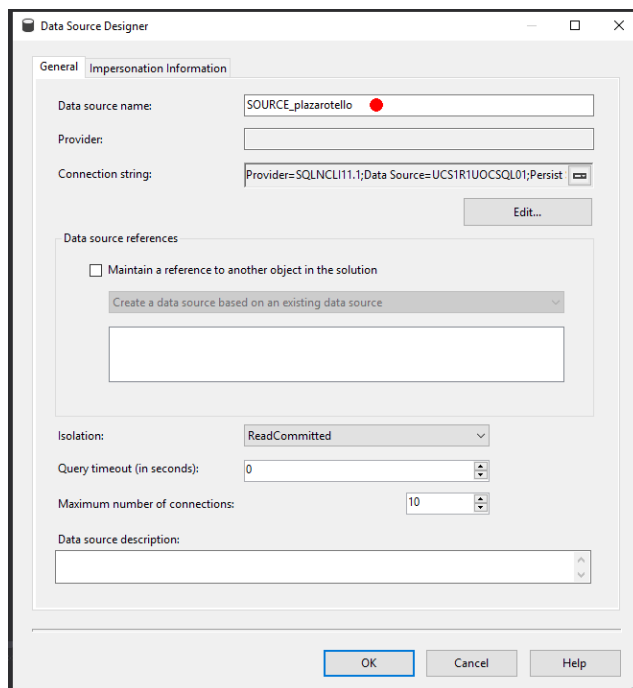
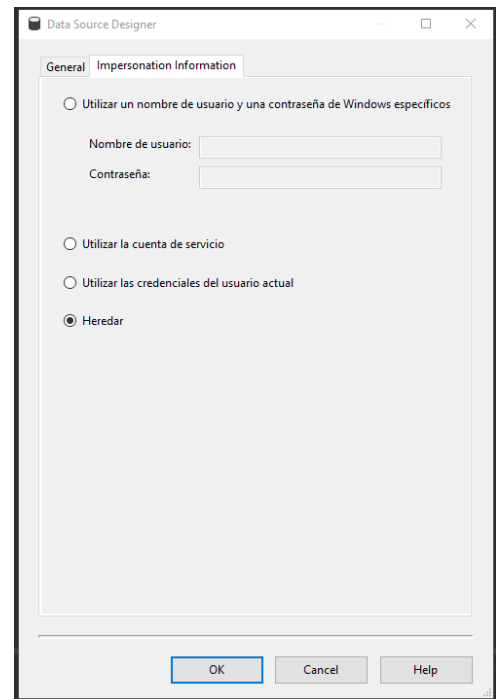


Figura 3: Comprobación de la conexión a la base de datos



The 'Data Source Designer' window shows the 'General' tab. The 'Data source name' is 'SOURCE_plazarotello'. The 'Provider' is empty. The 'Connection string' is 'Provider=SQLNCLI11.1;Data Source=UCS1R1UOCSQL01;Persist=1'. The 'Data source references' section has a checkbox 'Maintain a reference to another object in the solution' which is unchecked. Below it is a dropdown menu 'Create a data source based on an existing data source' and an empty text box. The 'Isolation' is set to 'ReadCommitted'. The 'Query timeout (in seconds)' is '0'. The 'Maximum number of connections' is '10'. The 'Data source description' is empty. At the bottom are 'OK', 'Cancel', and 'Help' buttons.

(a) Configuración de la conexión a la base de datos



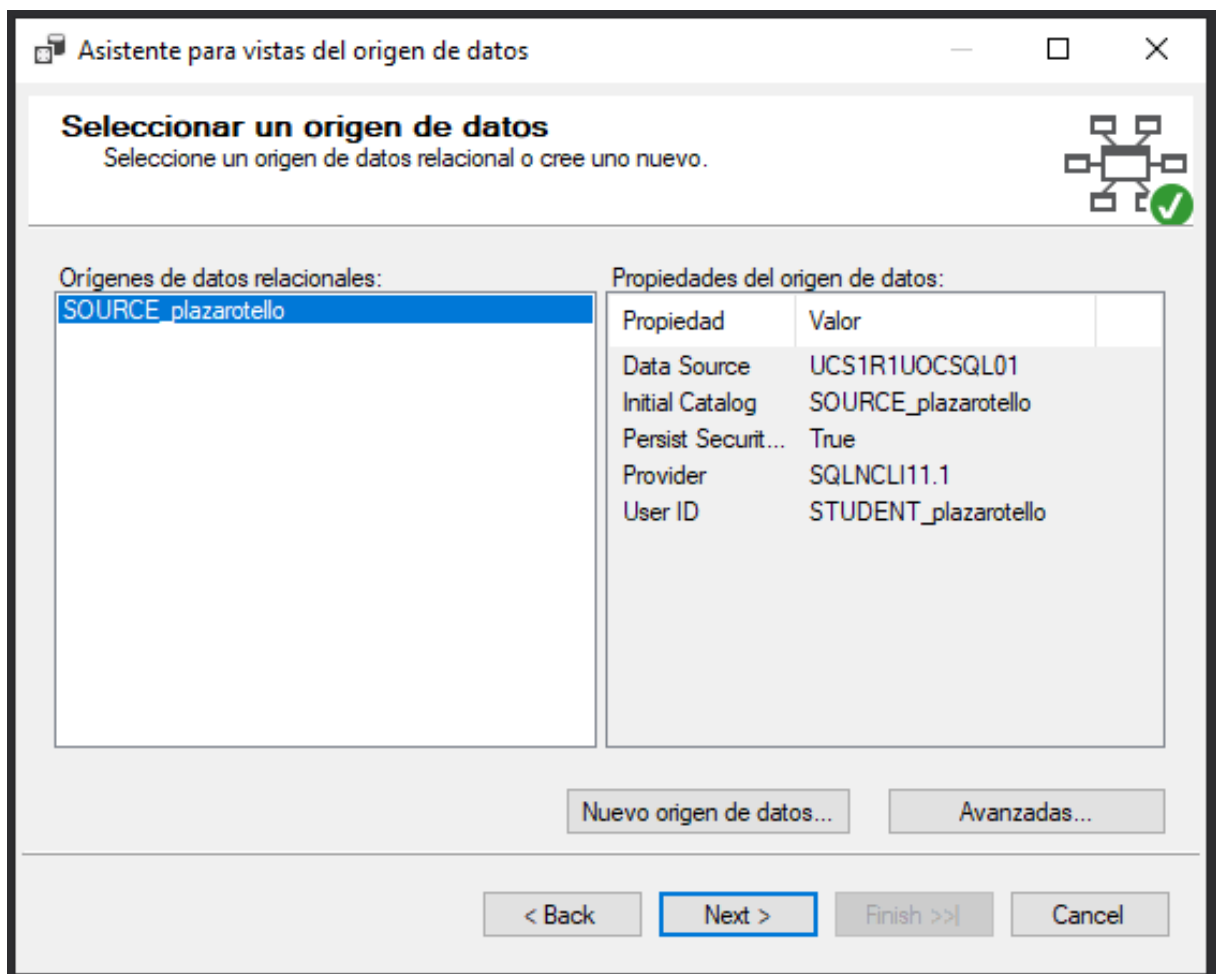
The 'Data Source Designer' window shows the 'Impersonation Information' tab. It has three radio buttons: 'Utilizar un nombre de usuario y una contraseña de Windows específicos', 'Utilizar la cuenta de servicio', and 'Utilizar las credenciales del usuario actual'. The first option is selected. Below the first option are fields for 'Nombre de usuario:' and 'Contraseña:'. Below the second option is a field for 'Nombre de usuario:'. Below the third option is a field for 'Nombre de usuario:'. At the bottom are 'OK', 'Cancel', and 'Help' buttons.

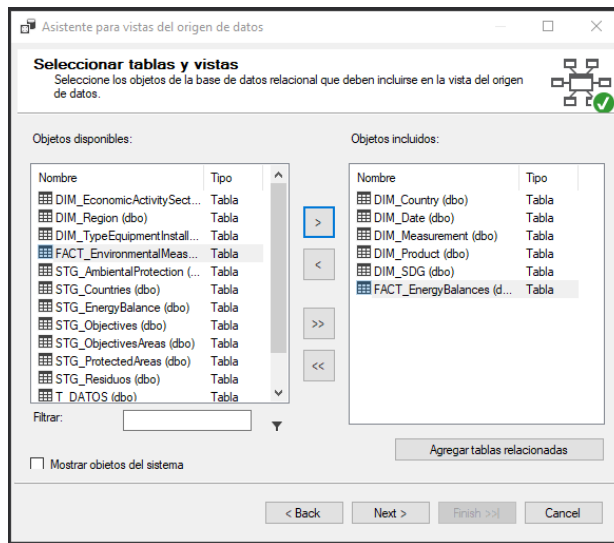
(b) Información de impersonalización

2.3. Vista del origen de datos

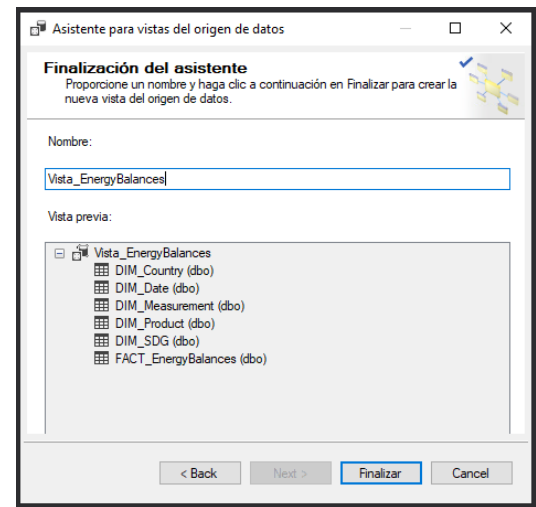
El siguiente paso consiste en la creación de las vistas de orígenes de datos. Tal y como se definió en el documento de análisis y diseño, y como se implementó en la fase de carga de datos, se procede a crear una vista por cada hecho en el modelo multidimensional.

En ambos casos se utiliza el mismo origen de datos, «SOURCE_plazarotello»:



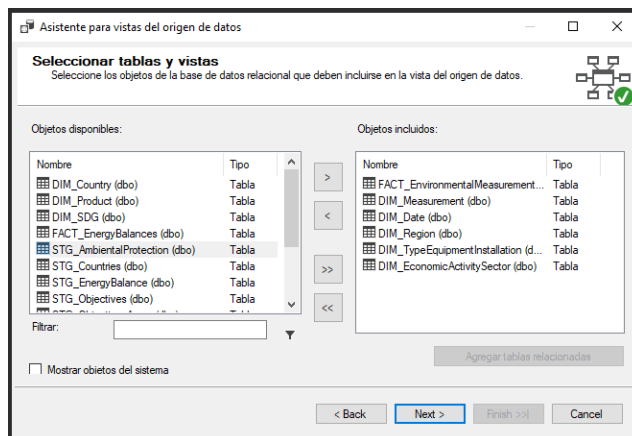


(c) Selección de tablas

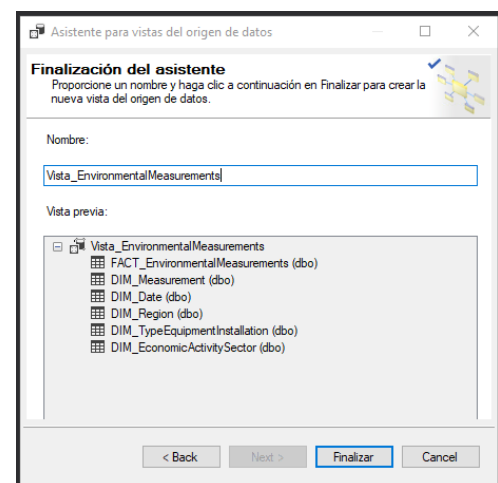


(d) Resumen de la vista

Figura 4: Creación de la vista para «FACT_EnergyBalances»



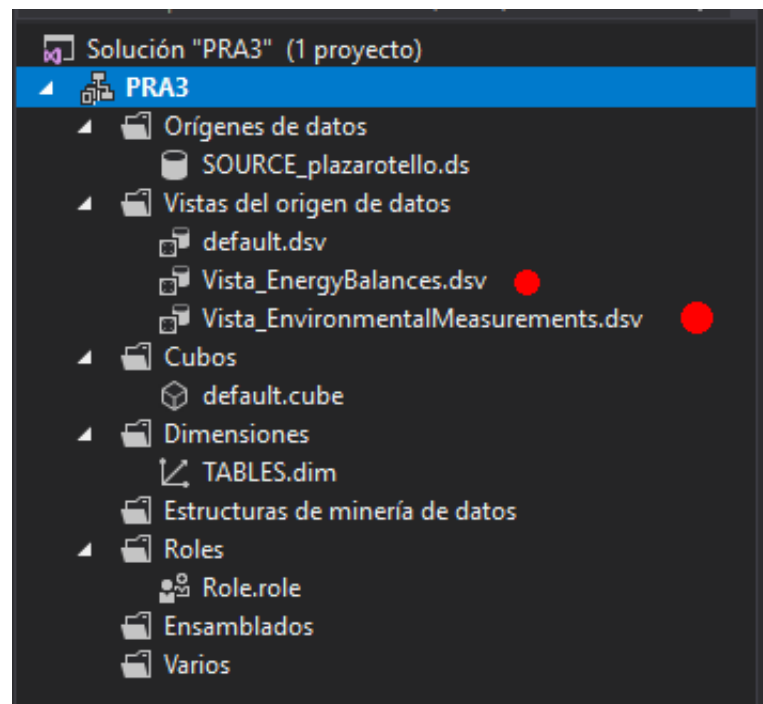
(a) Selección de tablas



(b) Resumen de la vista

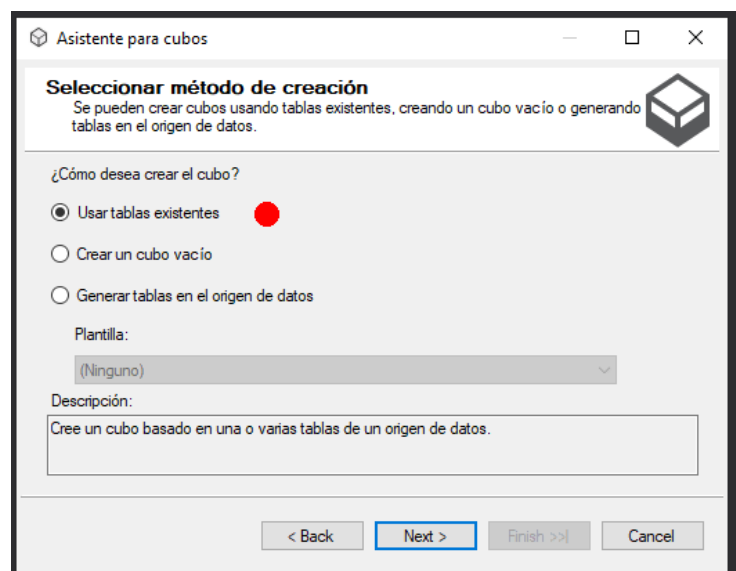
Figura 5: Creación de la vista para «FACT_EnergyBalances»

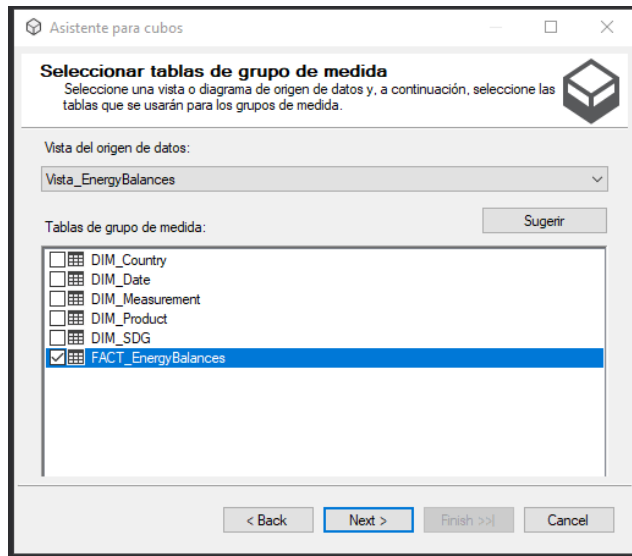
Tras la creación de la vistas en el proyecto, la estructura de la solución queda:



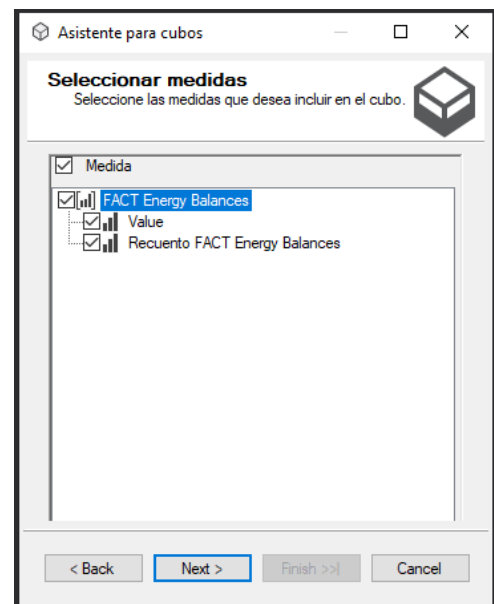
2.4. Creación e implementación de los cubos

Para cada vista creada en la sección anterior, se va a implementar el cubo correspondiente, configurando las tablas que son medidas y las tablas que son dimensiones. En ambos casos se utilizan las tablas existentes para crear el cubo:

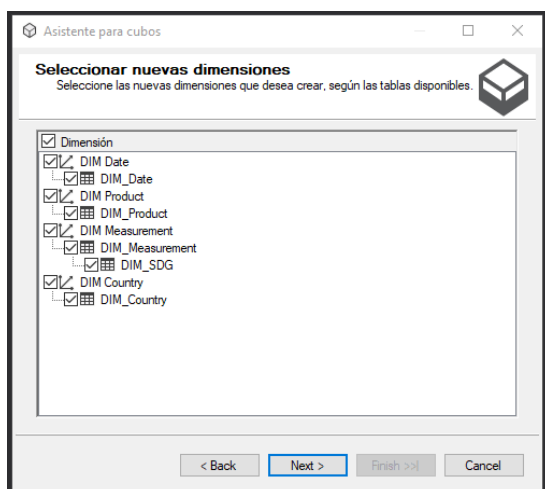




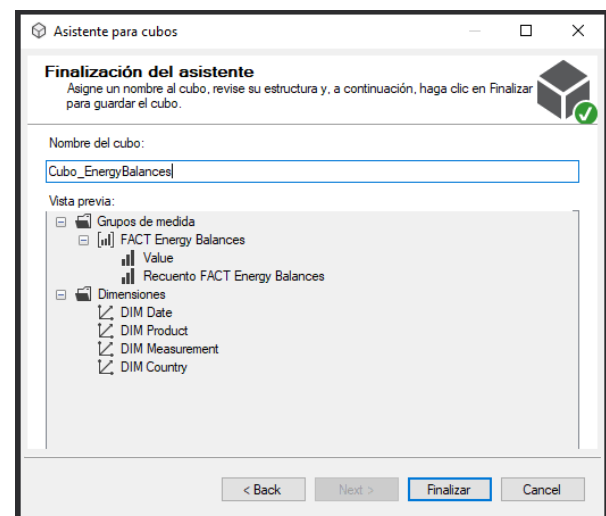
(a) Selección de la tabla de hechos



(b) Selección de las medidas



(c) Selección de las dimensiones



(d) Resumen

Figura 6: Creación del cubo para «FACT_EnergyBalances»

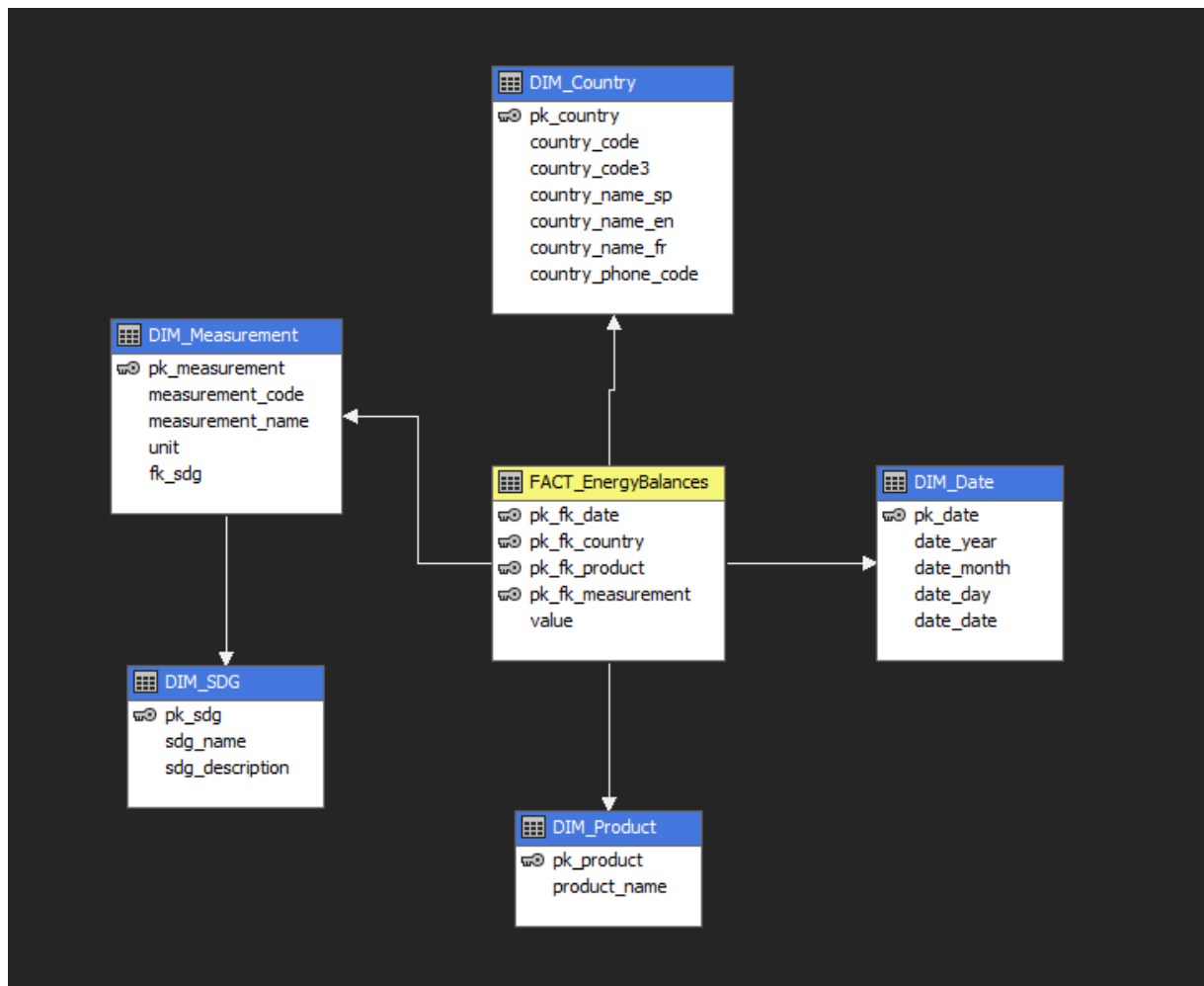
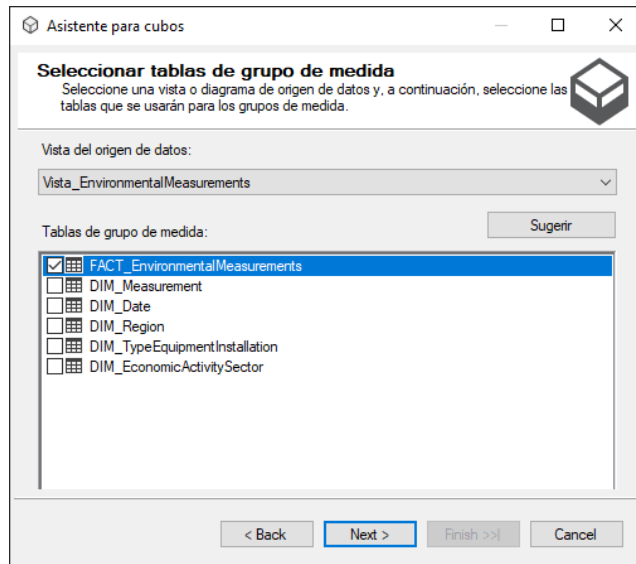
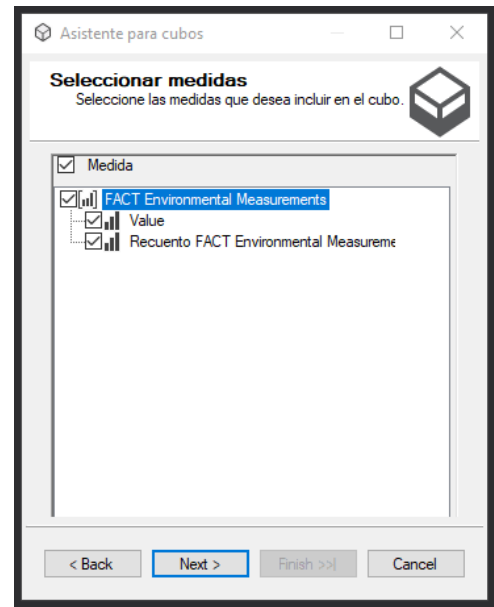


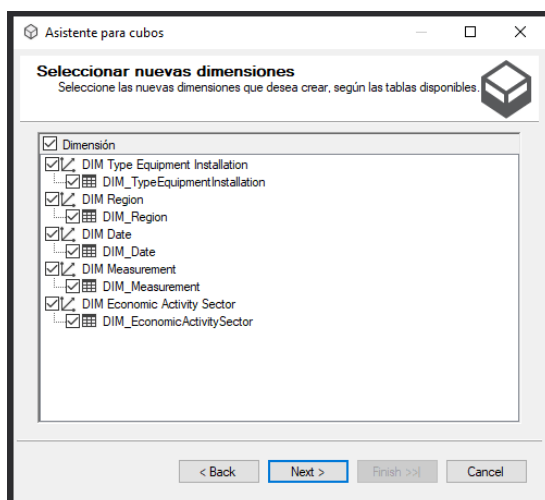
Figura 7: Estructura del cubo de «FACT_EnergyBalances»



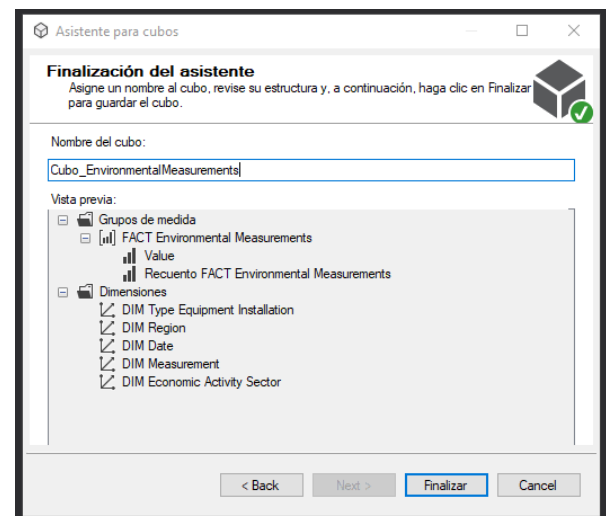
(a) Selección de la tabla de hechos



(b) Selección de las medidas



(c) Selección de las dimensiones



(d) Resumen

Figura 8: Creación del cubo para «FACT_EnvironmentalMeasurements»

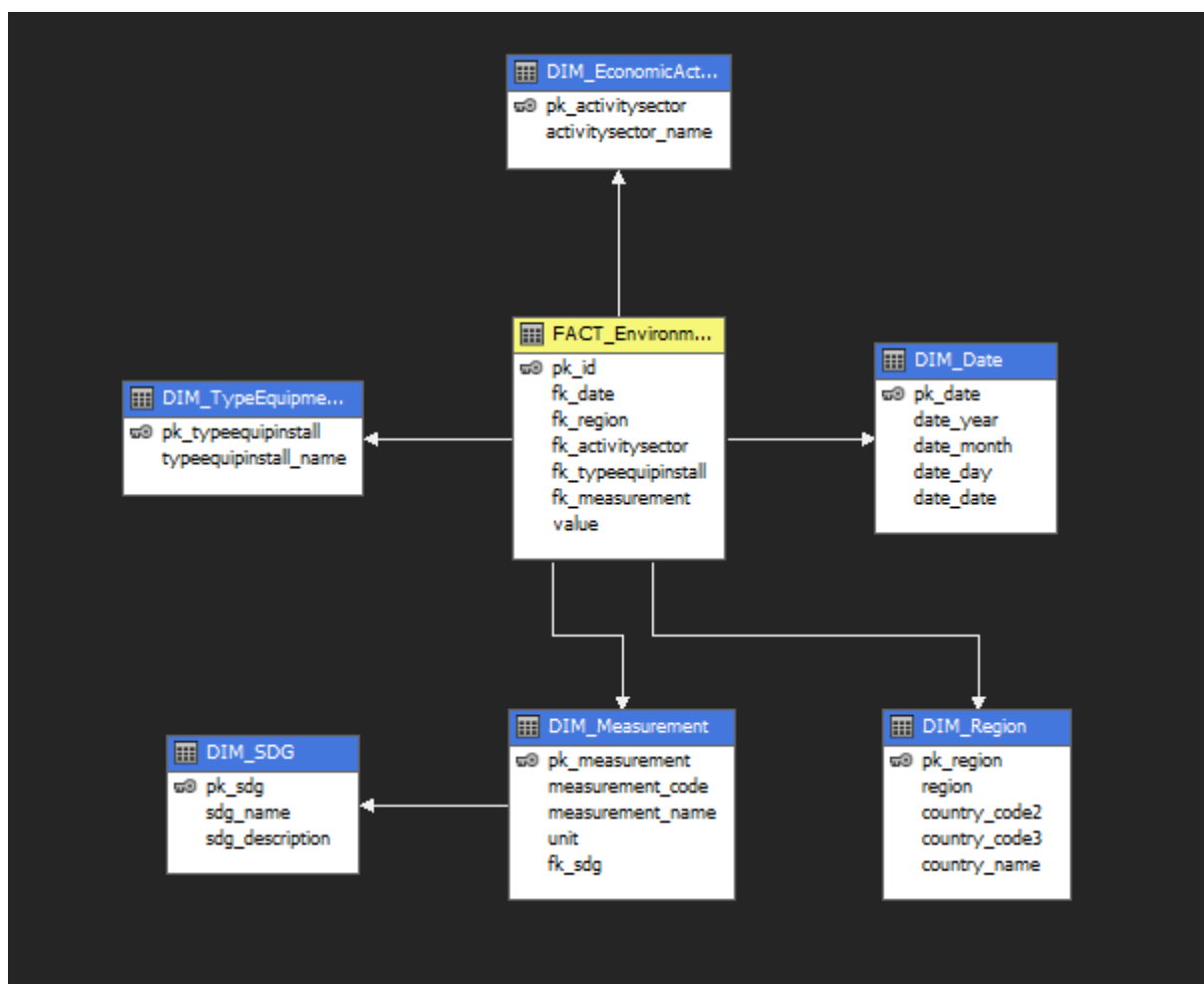


Figura 9: Estructura del cubo de «FACT_EnvironmentalMeasurements»

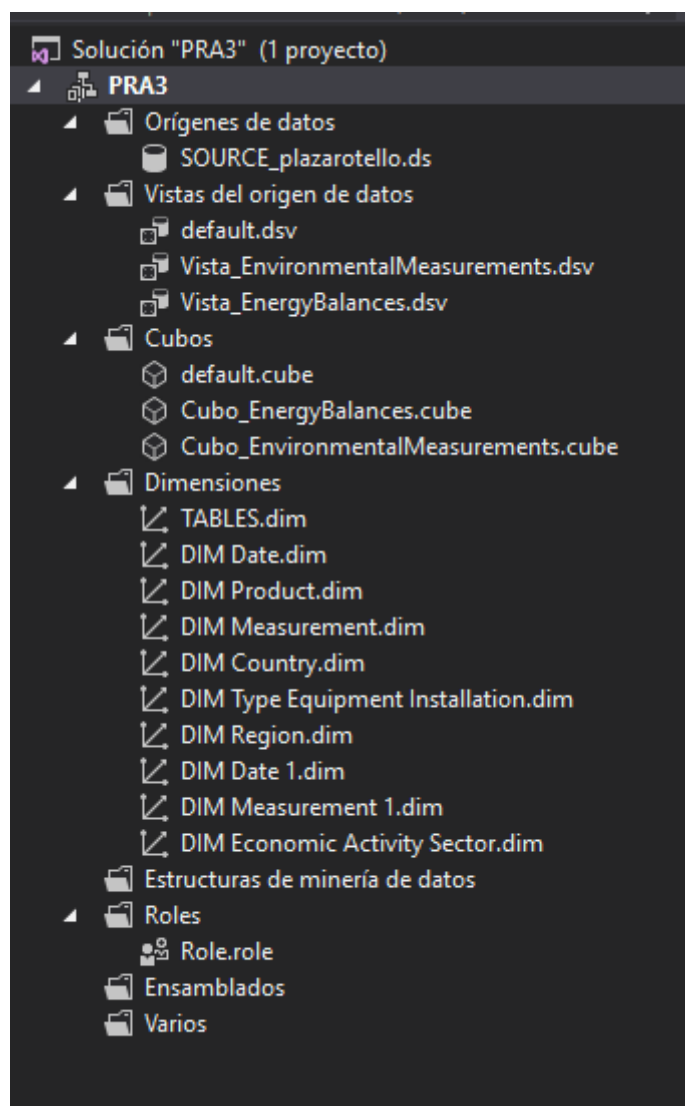
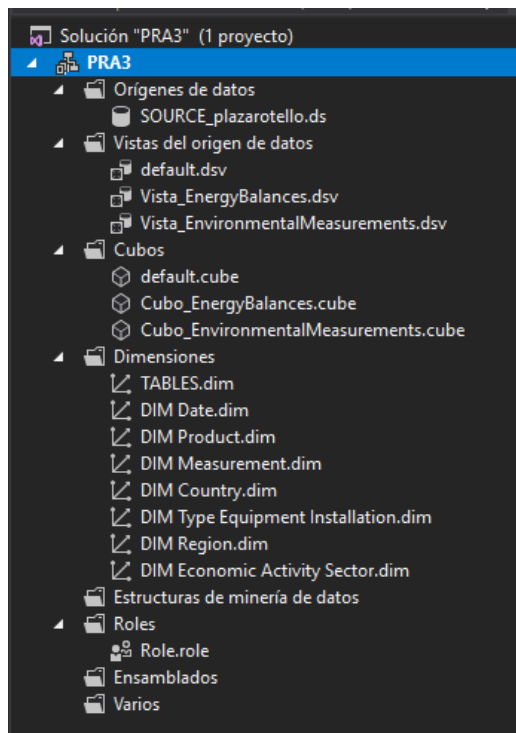


Figura 10: Estructura de la solución tras la creación de los cubos

En la sección de dimensiones se han creado las dimensiones que alberga cada cubo. Algunas de ellas, como «DIM_Date», están repetidas.

2.5. Configuración de dimensiones conformadas

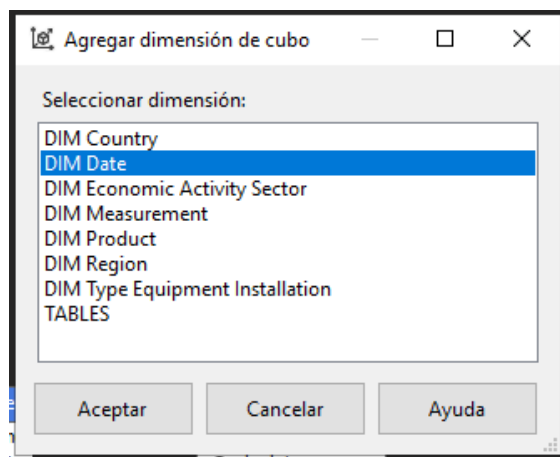
Se observan dimensiones duplicadas en la carpeta de dimensiones. Se procede en primer lugar a borrarlas, quedando la solución:



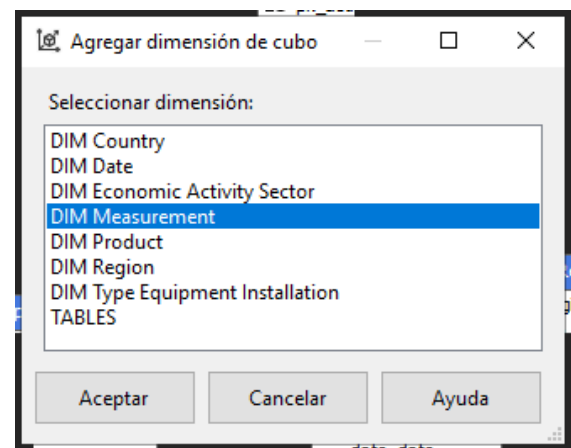
Aunque ahora ya no existen dimensiones duplicadas, es necesario reconfigurar los cubos para que apunten a las dimensiones correctas.

En este caso, las dimensiones duplicadas provenían del cubo de «FACT_EnvironmentalMeasurements», y es ese cubo el que debe repararse para apuntar a las dimensiones correctas.

Para solucionar estos problemas hay que, en primer lugar, agregar las dimensiones perdidas en el borrado al cubo:



(a) Dimensión de fecha




(b) Dimensión de medida

A continuación, se define la relación entre las dimensiones y la tabla de hechos en el cubo:

Definir relación

Seleccionar tipo de relación: Normal

La tabla de dimensiones está unida directamente a la tabla de hechos.



Atributo de granularidad: Pk Date

Tabla de dimensiones: DIM_Date

Tabla de grupos de medida: FACT_EnvironmentalMeasurements

Relación:

Columnas de dimensión	Columnas de grupo de medida
pk_date	fk_date

Avanzadas...


Aceptar Cancelar Ayuda

(c) Relación de la dimensión de fecha

Definir relación

Seleccionar tipo de relación: Normal

La tabla de dimensiones está unida directamente a la tabla de hechos.



Atributo de granularidad: Pk Measurement

Tabla de dimensiones: DIM_Measurement

Tabla de grupos de medida: FACT_EnvironmentalMeasurements

Relación:

Columnas de dimensión	Columnas de grupo de medida
pk_measurement	fk_measurement

Avanzadas...

Aceptar Cancelar Ayuda

(d) Relación de la dimensión de medida

Así, el cubo queda de nuevo
configurado correctamente sin
dimensiones duplicadas:

Grupos de medida	
Dimensiones	FACT Environmental Measur...
DIM Type Equipment Install...	Pk Typeequipinstall
DIM Region	Pk Region
DIM Economic Activity Sector	Pk Activitysector
DIM Measurement	Pk Measurement
DIM Date	Pk Date

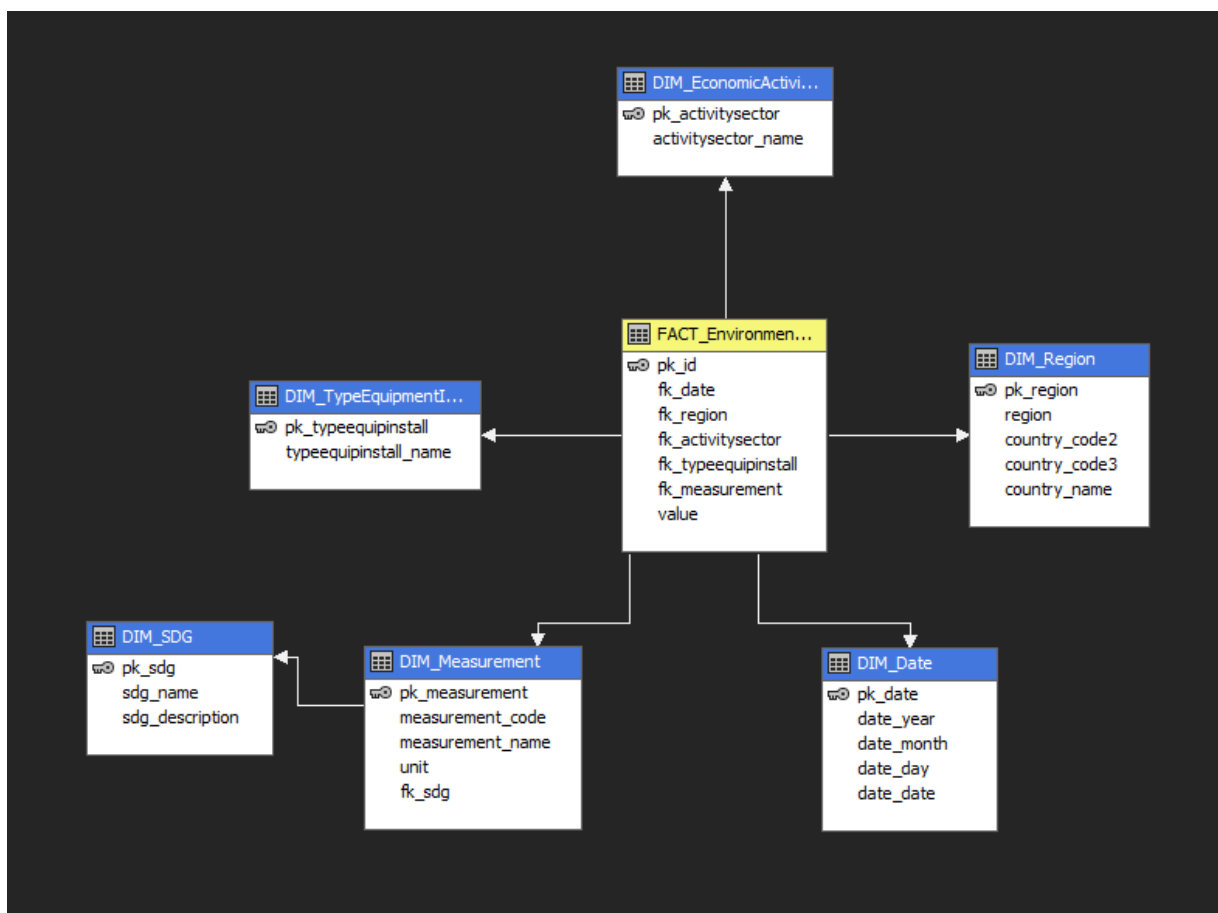
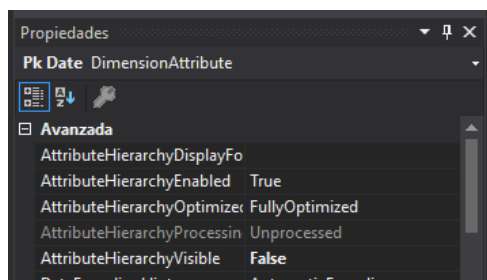


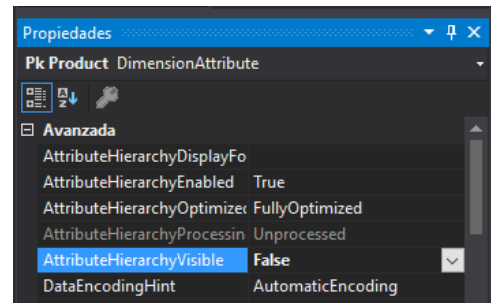
Figura 11: Estructura del cubo

2.6. Jerarquías y dimensiones

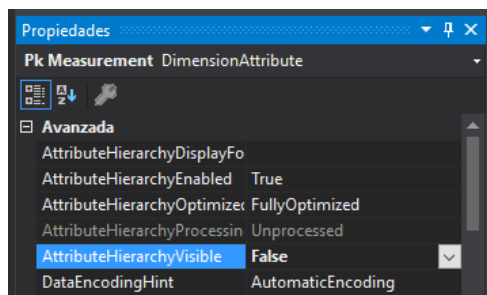
En primer lugar se procede a ocultar las claves primarias de las dimensiones. Aunque son útiles para el sistema de base de datos, no aportan más que confusión a la hora de realizar consultas mediante una herramienta de este estilo.



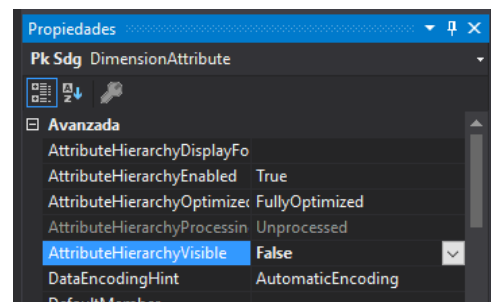
(a) Pk_Date



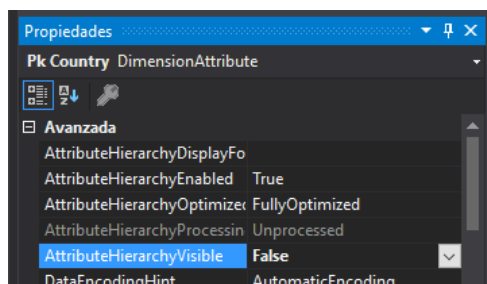
(b) Pk_Product



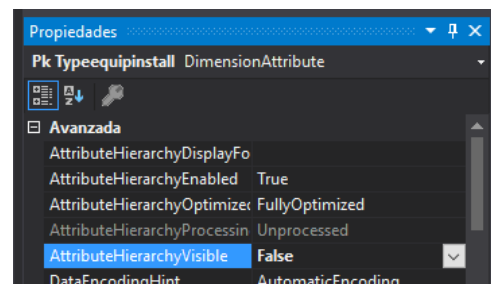
(c) Pk_Measurement



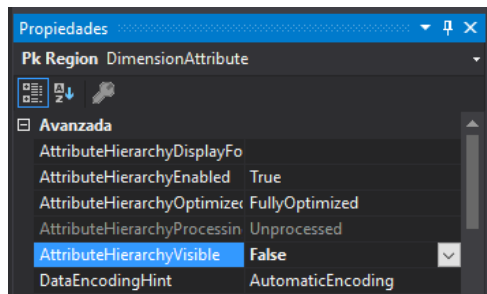
(d) Pk_Sdg



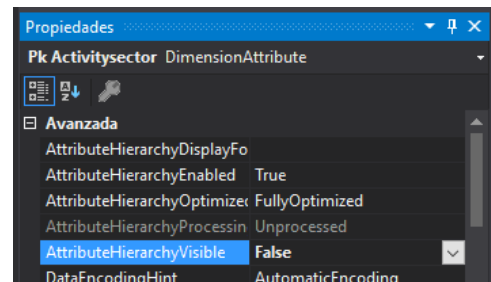
(e) Pk_Country



(f) Pk_TypeEquipInstall

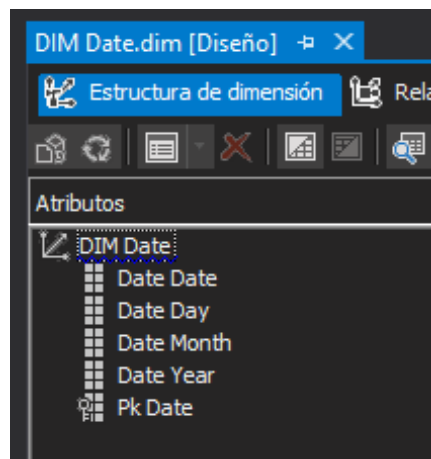


(g) Pk_Region

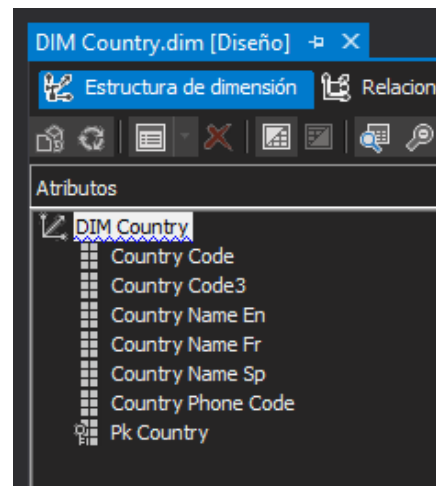


(h) Pk_ActivitySector

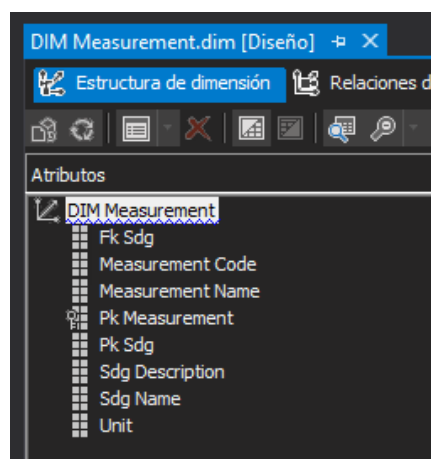
A continuación se añaden los atributos que se van a utilizar para las consultas:



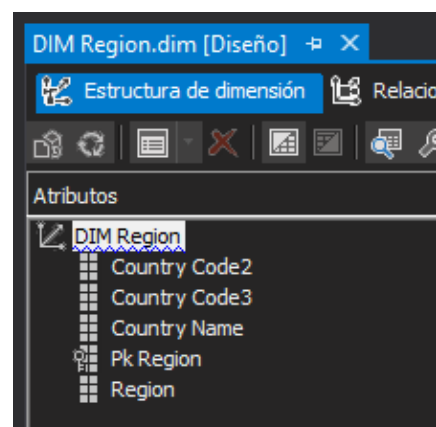
(i) DIM_Date



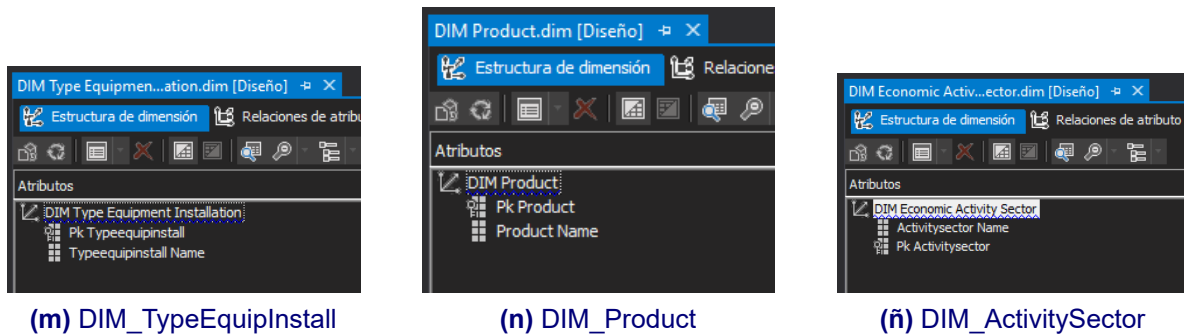
(j) DIM_Country



(k) DIM_Measurement



(l) DIM_Region



Finalmente se ha creado una jerarquía para la dimensión temporal y para la dimensión Región:

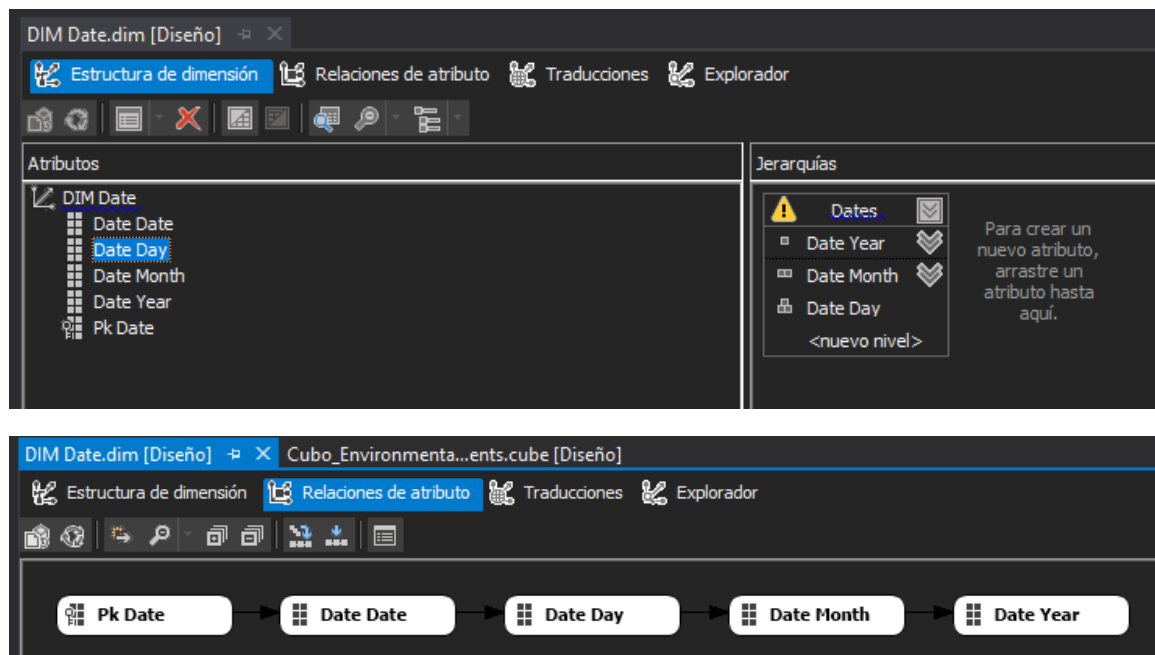


Figura 12: Jerarquía para la dimensión temporal «DIM_Date»

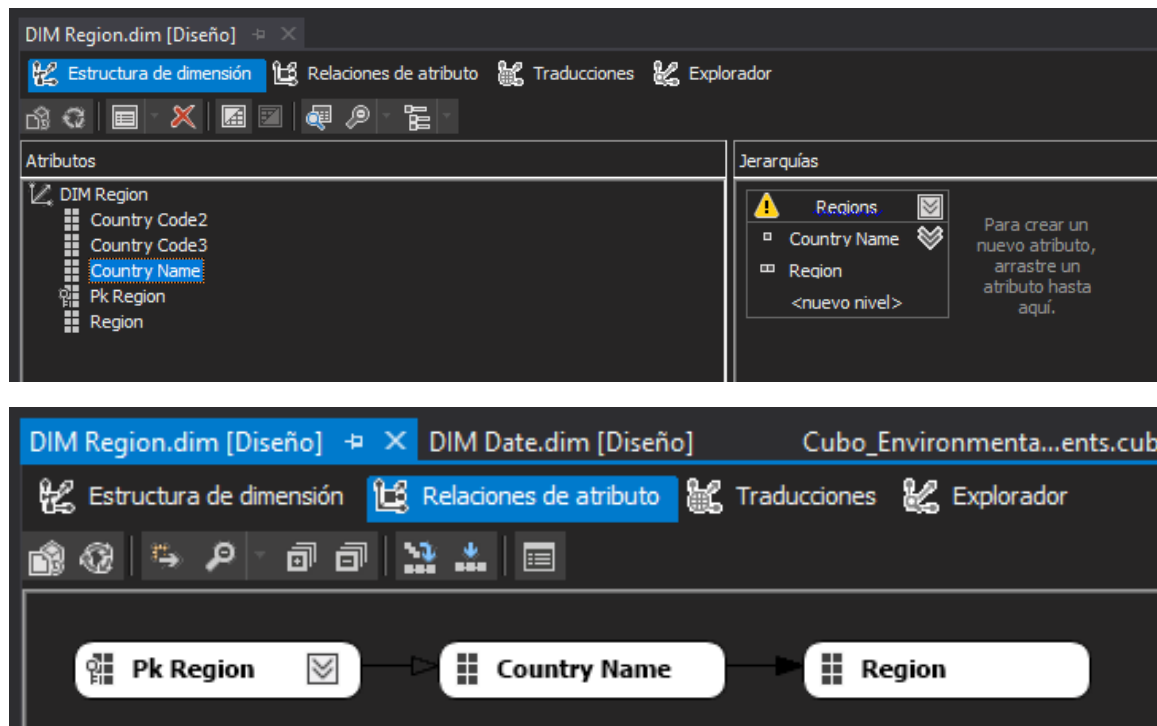


Figura 13: Jerarquía para la dimensión «DIM_Region»

2.7. Procesado y resolución de errores

El último paso en la creación del modelo OLAP consiste en implementar la solución, procesar dimensiones y cubos, y resolver posibles errores que puedan surgir. Al implementar la solución se obtiene:

Lista de errores						
Toda la solución						
		0 Errores	10 Advertencias	0 Mensajes	Compilación + IntelliSense	Lista de errores de búsqueda
	Códi...	Descripción	Proyecto	Archivo	Lí...	Estado suprimido
⚠		Dimension [DIM Economic Activity Sector]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Dimension [DIM Measurement]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Dimension [DIM Type Equipment Installation]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Dimension [DIM Date]: Evite las jerarquías de atributos visibles para los atributos empleados como niveles en las jerarquías definidas por el usuario.			0	
⚠		Hierarchy [DIM Date].[Dates]: No existen relaciones de atributo entre uno o varios niveles de esta jerarquía, lo cual puede disminuir el rendimiento de las consultas.			0	
⚠		Dimension [DIM Country]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Dimension [DIM Product]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Cube [default]: Evite los cubos con una única dimensión.			0	
⚠		Dimension [DIM Region]: Cree jerarquías en dimensiones que no sean de elementos primarios y secundarios.			0	
⚠		Database [DEST_plazarotello]: La base de datos no tiene dimensión de tiempo. Considere la posibilidad de crearla.			0	

Figura 14: No hay errores, pero sí *warnings*

- Varios de estos errores se refieren a dimensiones sin jerarquía. Se decide ignorar este tipo de errores, ya que no todas las dimensiones tienen jerarquía posible.

- Aparece un error de que no existe dimensión de tiempo en la base de datos. Se procede a arreglarlo de forma sencilla: cambiando los tipos de los atributos de la dimensión «DIM_Date».

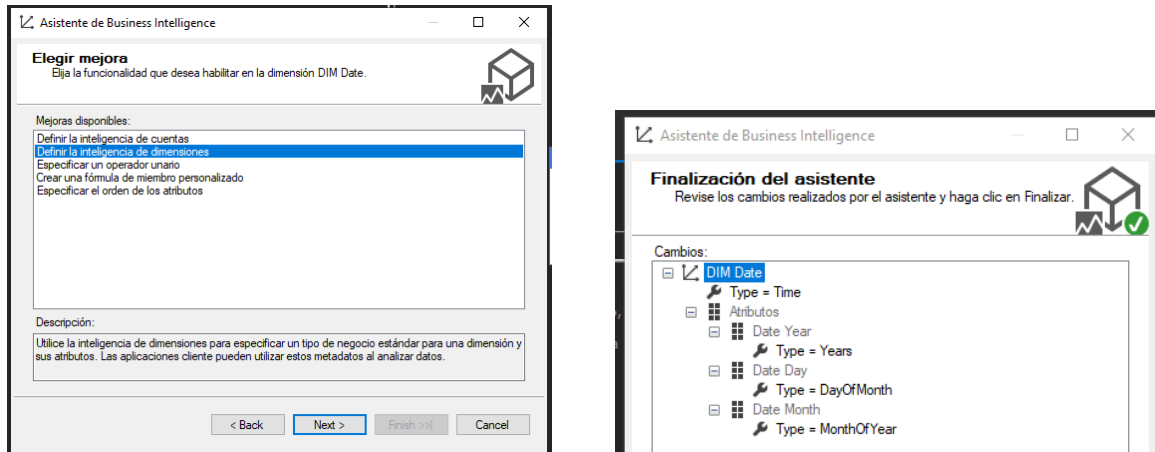


Figura 15: Uso del asistente de Business Intelligence para cambiar el tipo de los atributos

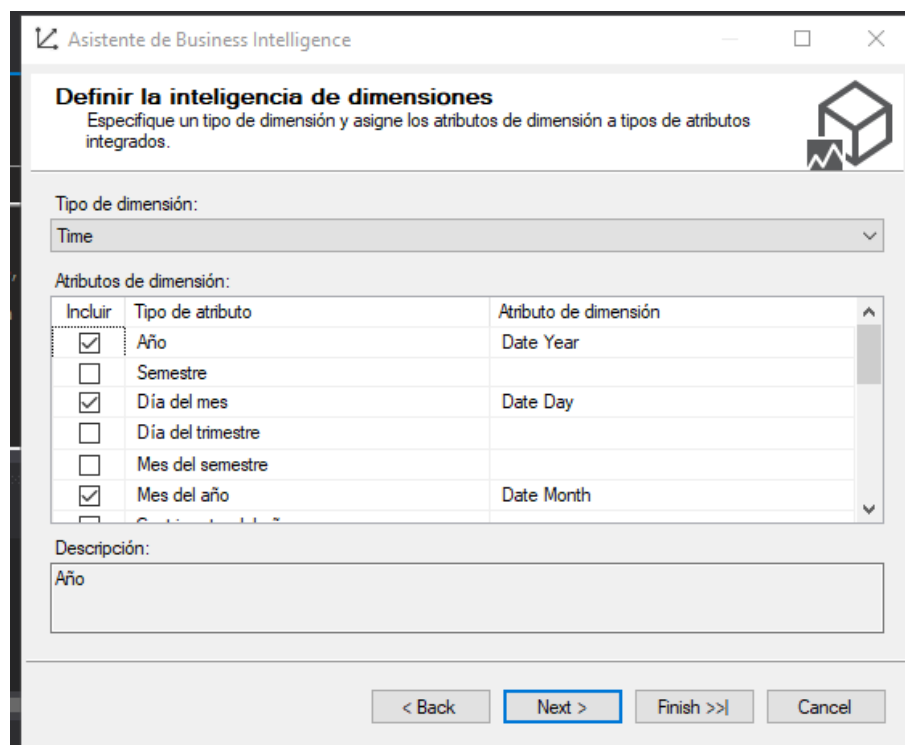
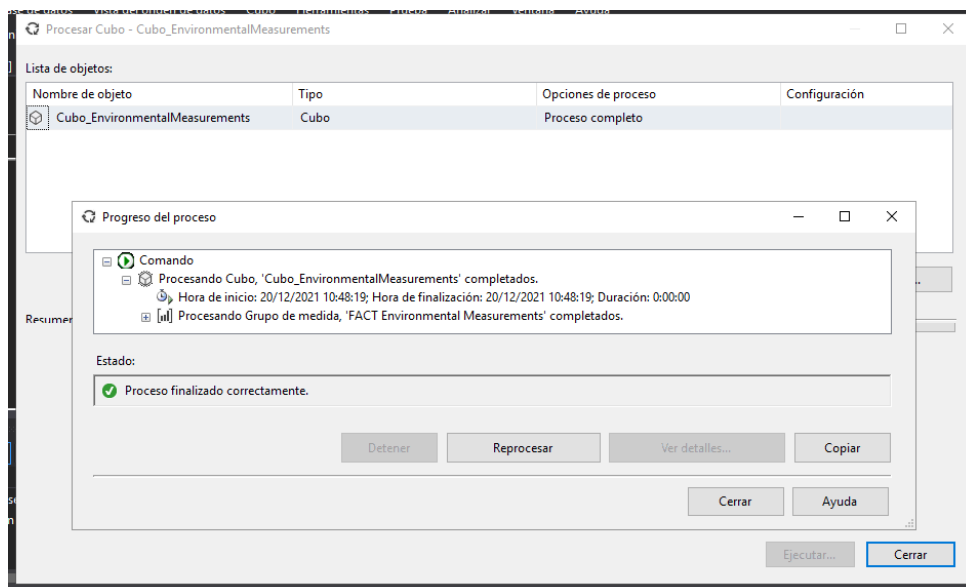
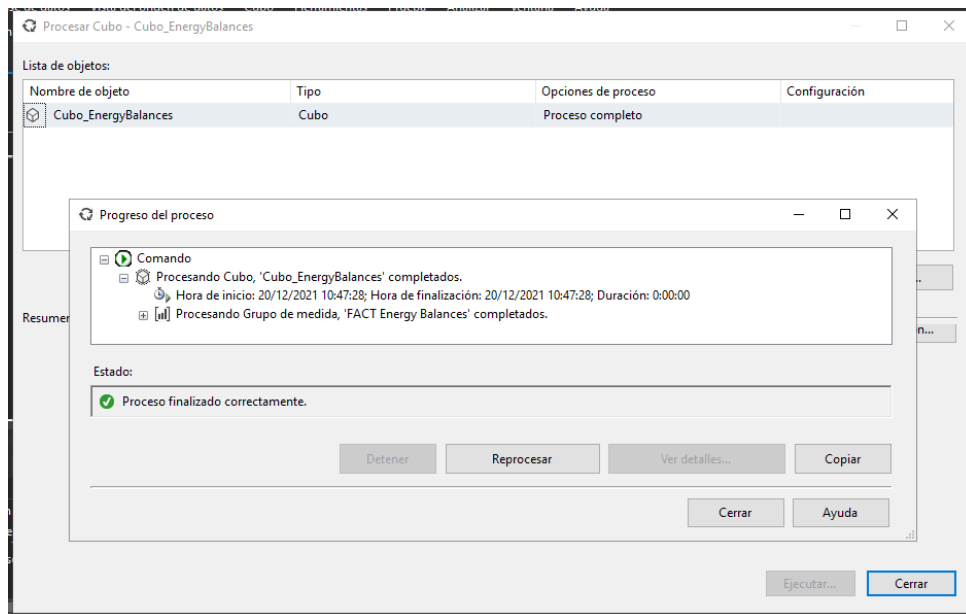
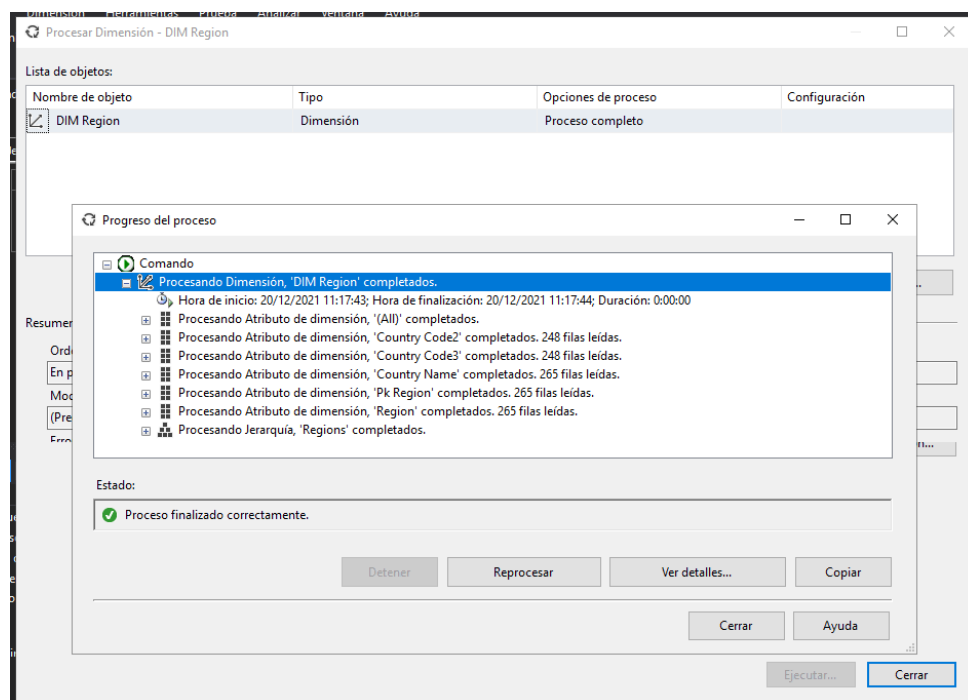
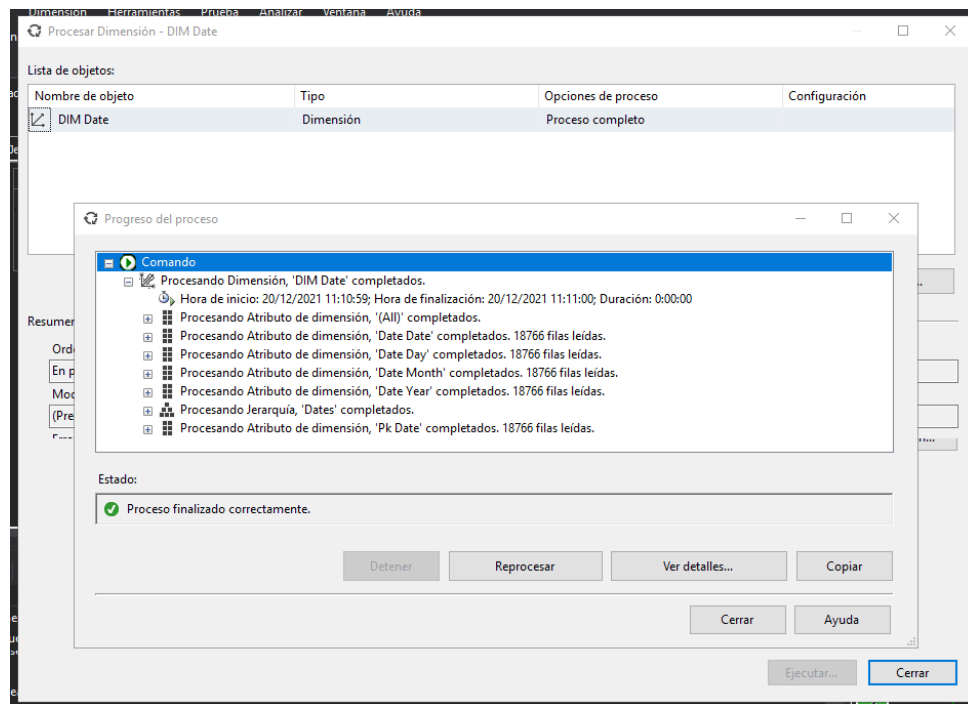


Figura 16: Los nuevos tipos de los atributos

- También aparece error porque algunos atributos participan en la jerarquía y están visibles fuera de la misma; se llevarán a cabo acciones similares que con las *primary keys* de las dimensiones.
- Por último, aparece un *warning* del cubo por defecto; se ignora.

A continuación se presentan los resultados del procesamiento de los cubos y las dimensiones con jerarquía:

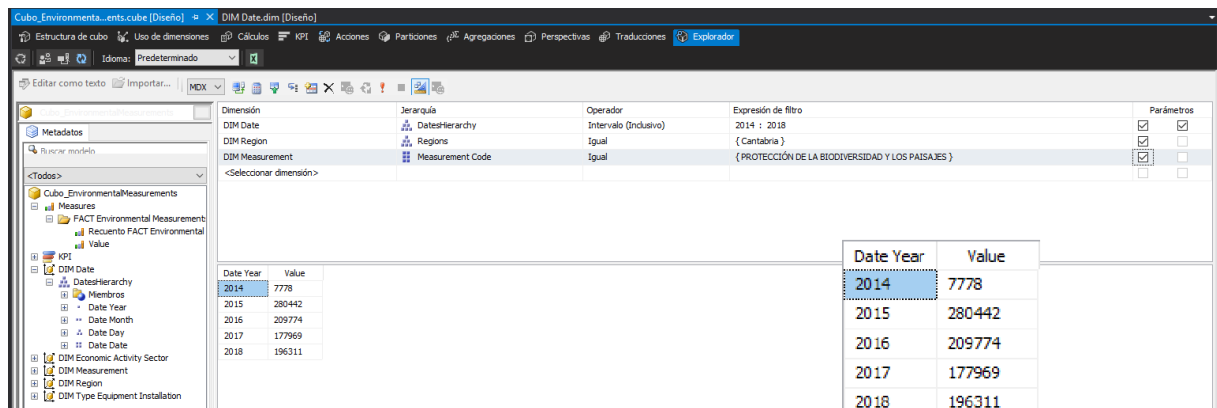




3. Explotación del modelo OLAP

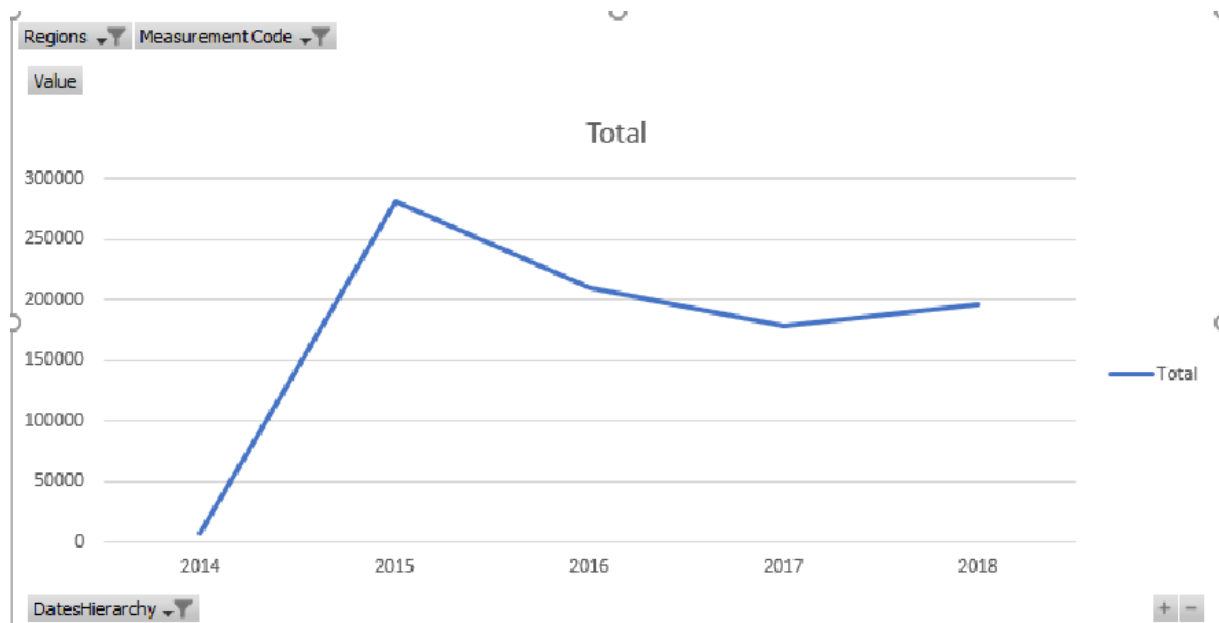
3.1. Análisis evolutivo de la inversión en protección de biodiversidad y paisajes en Cantabria (2014-2018)

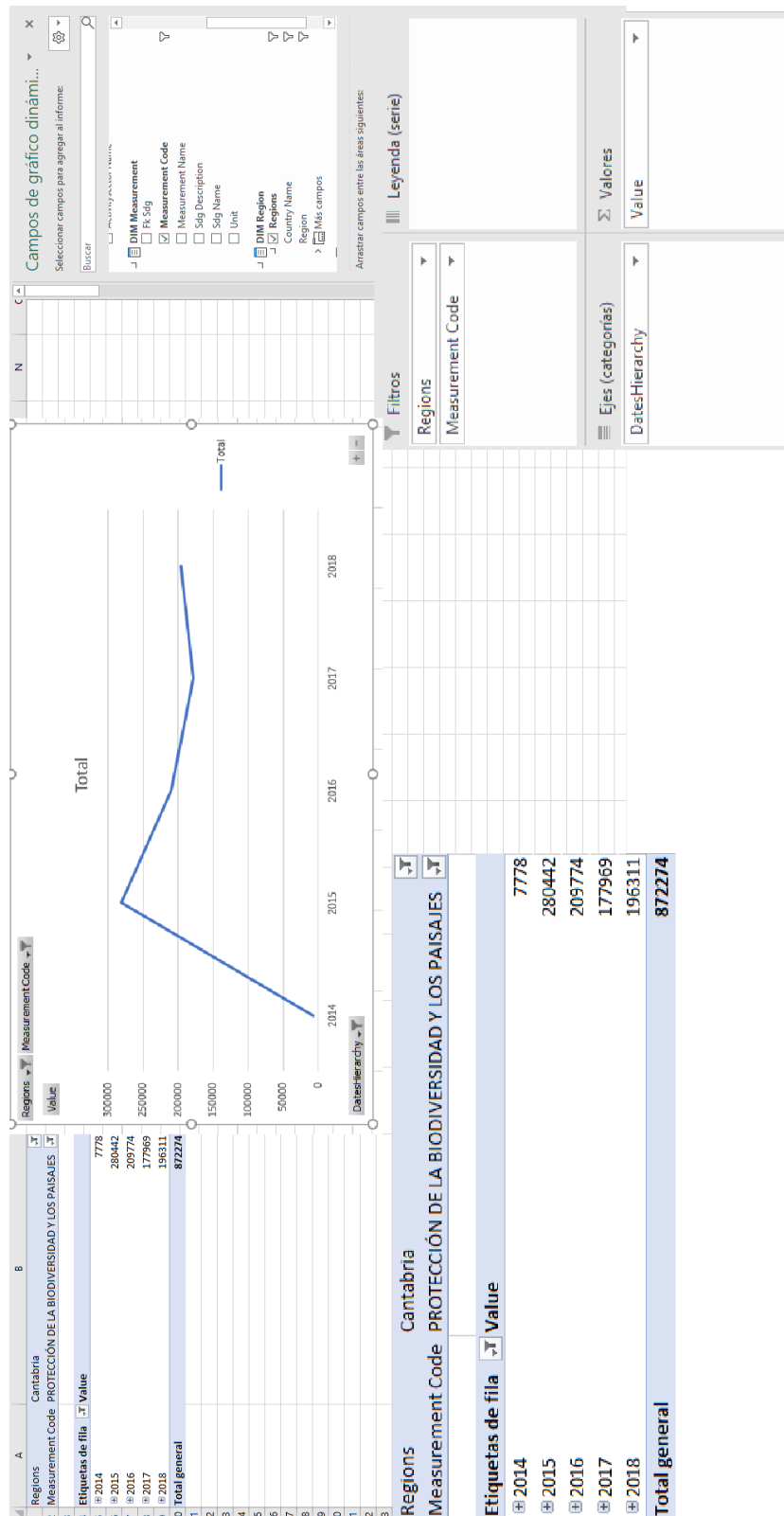
Esta consulta se ha realizado tanto en MDX como en Excel. Se ha utilizado el cubo de medidas ambientales y se ha filtrado por los años 2014 a 2018, la región de Cantabria y la medida de biodiversidad y paisajes. Por último, se ha añadido a los datos a visualizar tanto el valor resultante (*Value*) como el año al que hace referencia (*Date Year*).



Dimensión	Jerarquía	Operador	Expresión de filtro	Parámetros
DIM Date	DatesHierarchy	Intervalo (Inclusivo)	2014 : 2018	<input checked="" type="checkbox"/>
DIM Region	Regions	Igual	{ Cantabria }	<input checked="" type="checkbox"/>
DIM Measurement	Measurement Code	Igual	{ PROTECCIÓN DE LA BIODIVERSIDAD Y LOS PAISAJES }	<input checked="" type="checkbox"/>

Date Year	Value
2014	7778
2015	280442
2016	209774
2017	177969
2018	196311





A partir del año 2014 la inversión en biodiversidad y paisajes aumentó de forma considerable.

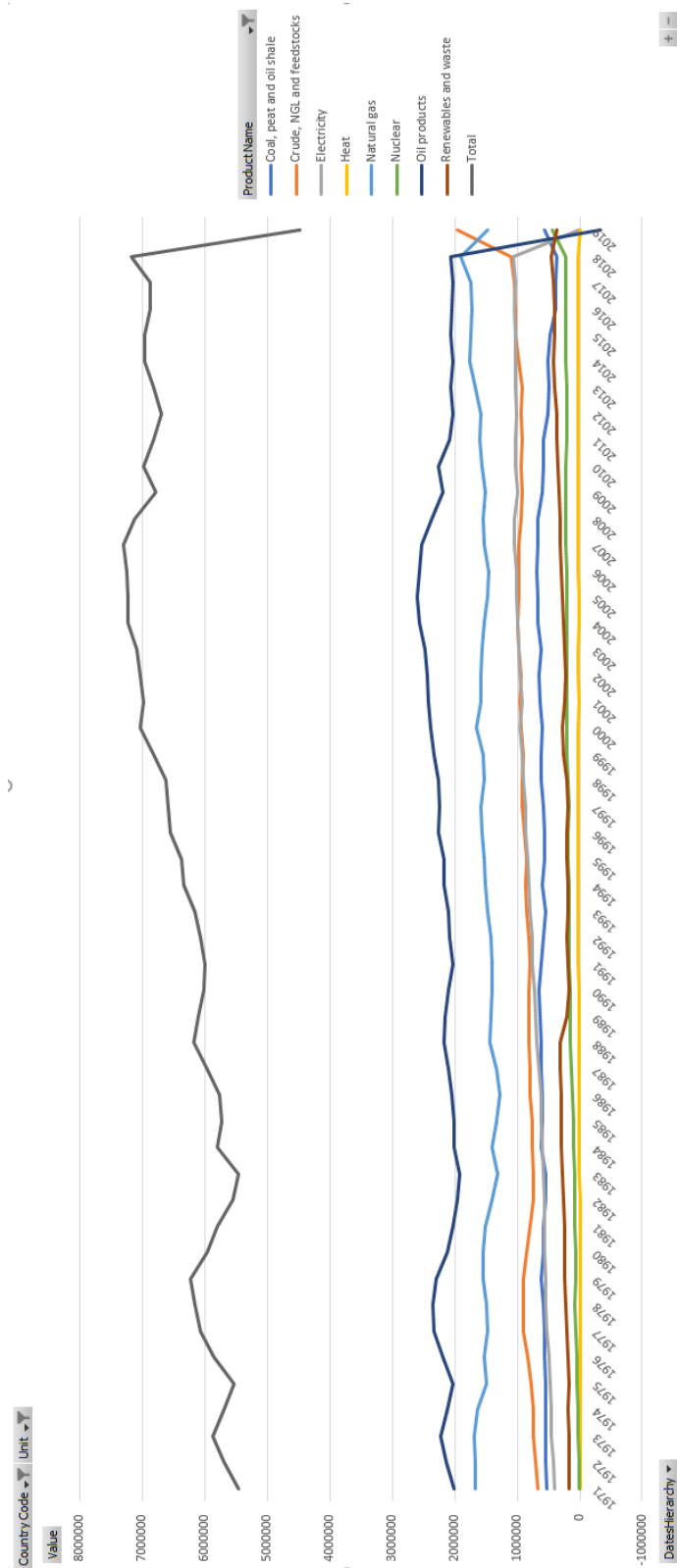
3.2. Análisis evolutivo de la producción en *ktoe* por producto de Estados Unidos

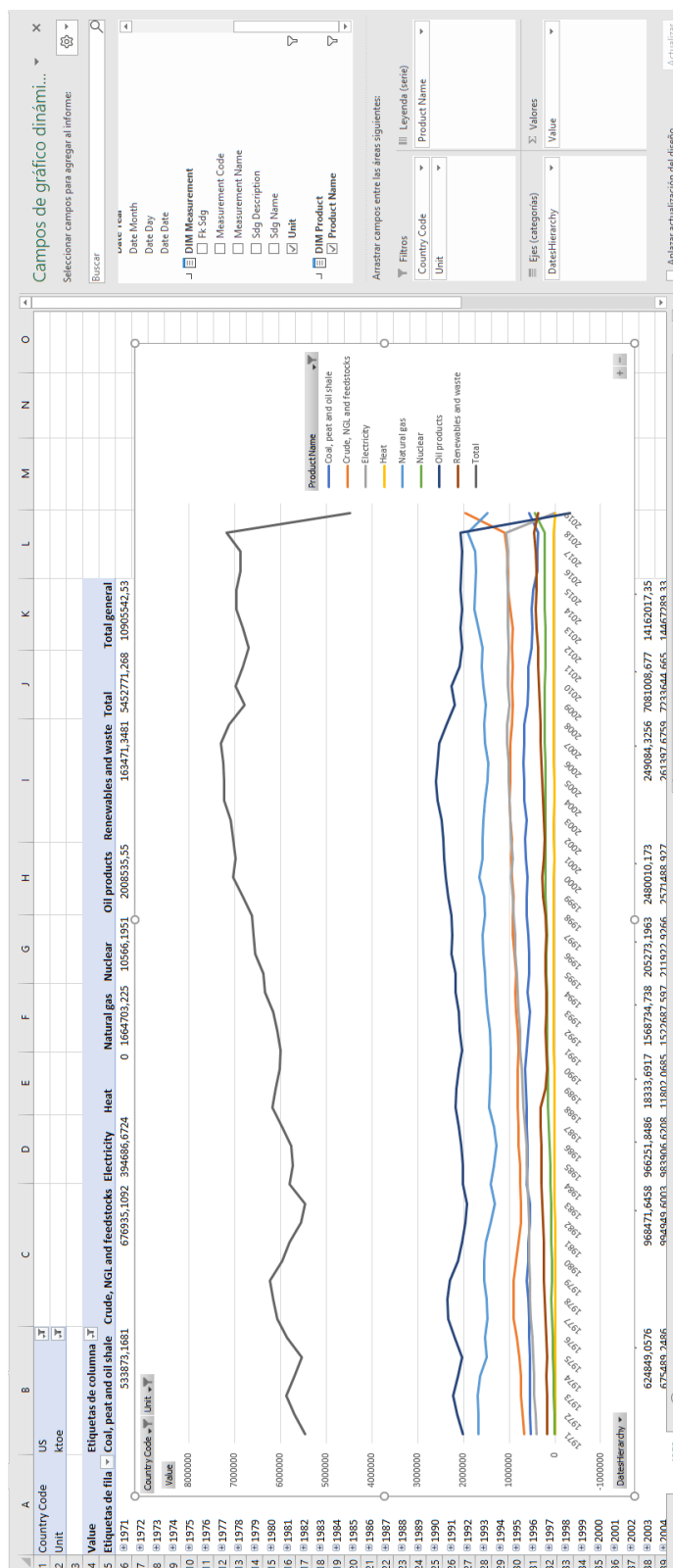
Se ha realizado la consulta en Excel, utilizando el cubo de balances energéticos. Se ha filtrado por el país (Estados Unidos) y las unidades (*ktoe*). Para mostrar los resultados en un gráfico dinámico, se han añadido los años al eje X.

Se muestra además el total del balance energético a lo largo de los años. Se observa una caída en la producción de gasolina en el último año en que se tienen datos (2019), que pasa de ser la energía más producida a la menos producida.

Caen también el gas natural y la electricidad en 2019. El petróleo sube de forma considerable, posicionándose como la energía más producida en Estados Unidos en 2019.

Del análisis evolutivo se obtienen, sin embargo, resultados confusos: no hay ningún año en que la variación de energía producida en el país haya sido tan fuerte como el último año, 2019. Se habría de investigar la razón de esta variación, si es debida a falta de datos o algún fenómeno que produjera la escasez.





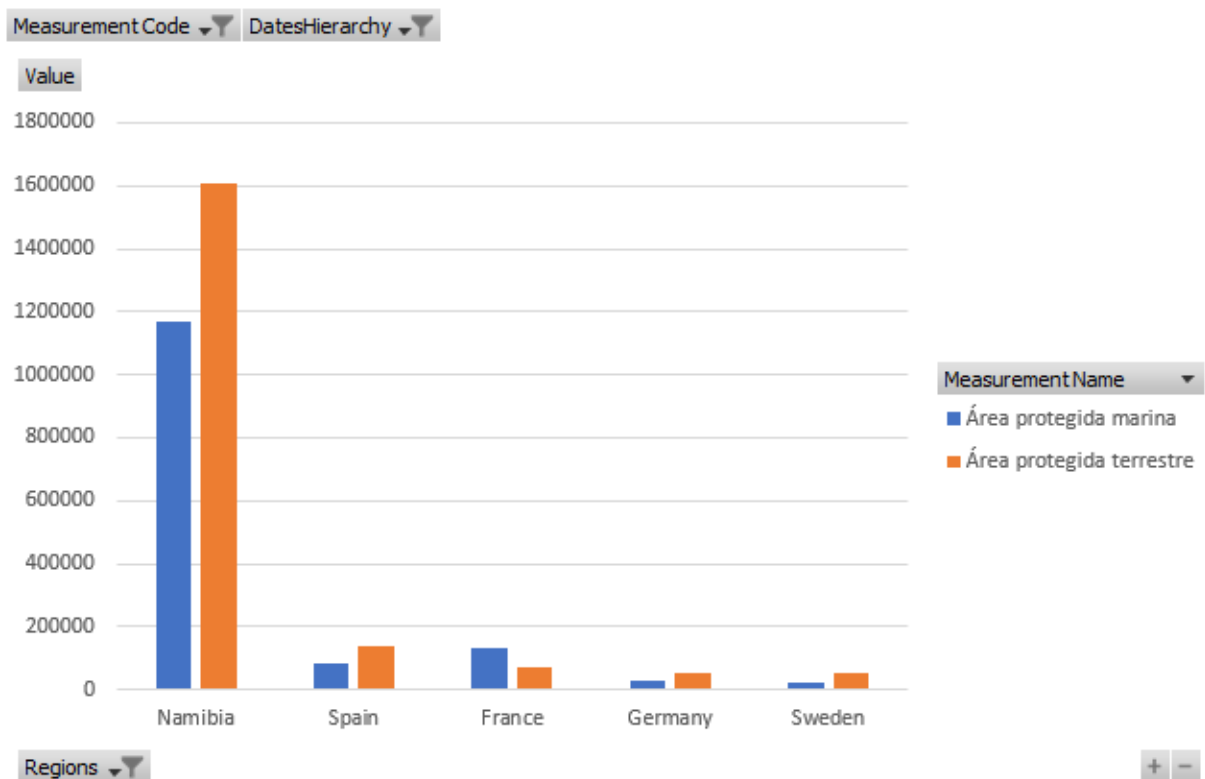
3.3. Análisis del *top five* de países con áreas protegidas mayores, tanto marinas como terrestres

Para este análisis se han realizado 3 consultas: el top 5 de países con áreas marinas protegidas mayores, el top 5 de países con áreas terrestres protegidas mayores y el top 5 de países con área protegida total mayor.

Se ha utilizado Excel y el cubo sobre el que se han desarrollado las consultas ha sido el de mediciones ambientales. Se ha filtrado por los códigos de medida MPA_KM2 y TPA_KM2, aunque también se podría haber filtrado por el nombre de la medida, disponible en la leyenda del gráfico.

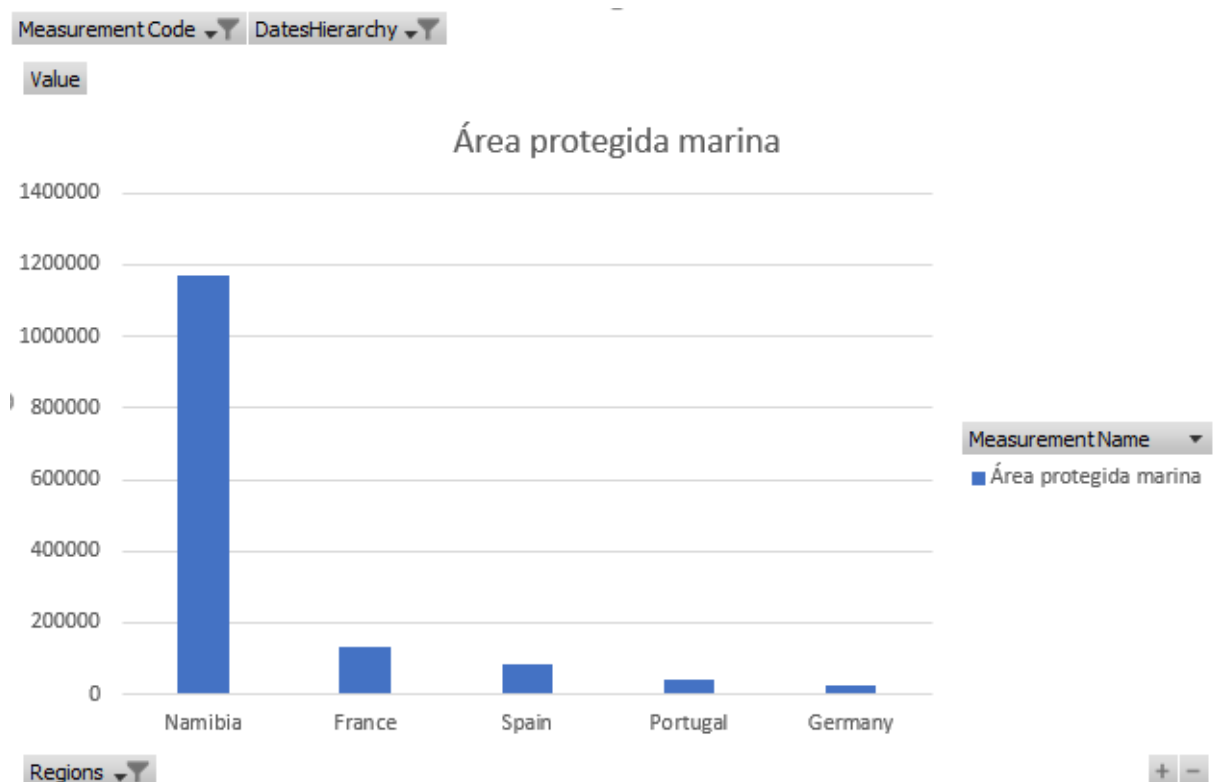
También se ha filtrado por fecha, utilizando solo el último año disponible (2019) para agregar las mediciones. Este análisis es necesario llevarlo a cabo tomando en cuenta la dimensión temporal; ya sea mediante un análisis evolutivo o fijando una fecha en que realizarlo.

Por último, se han ordenado las etiquetas de fila por el top 5 en descendente según el valor.

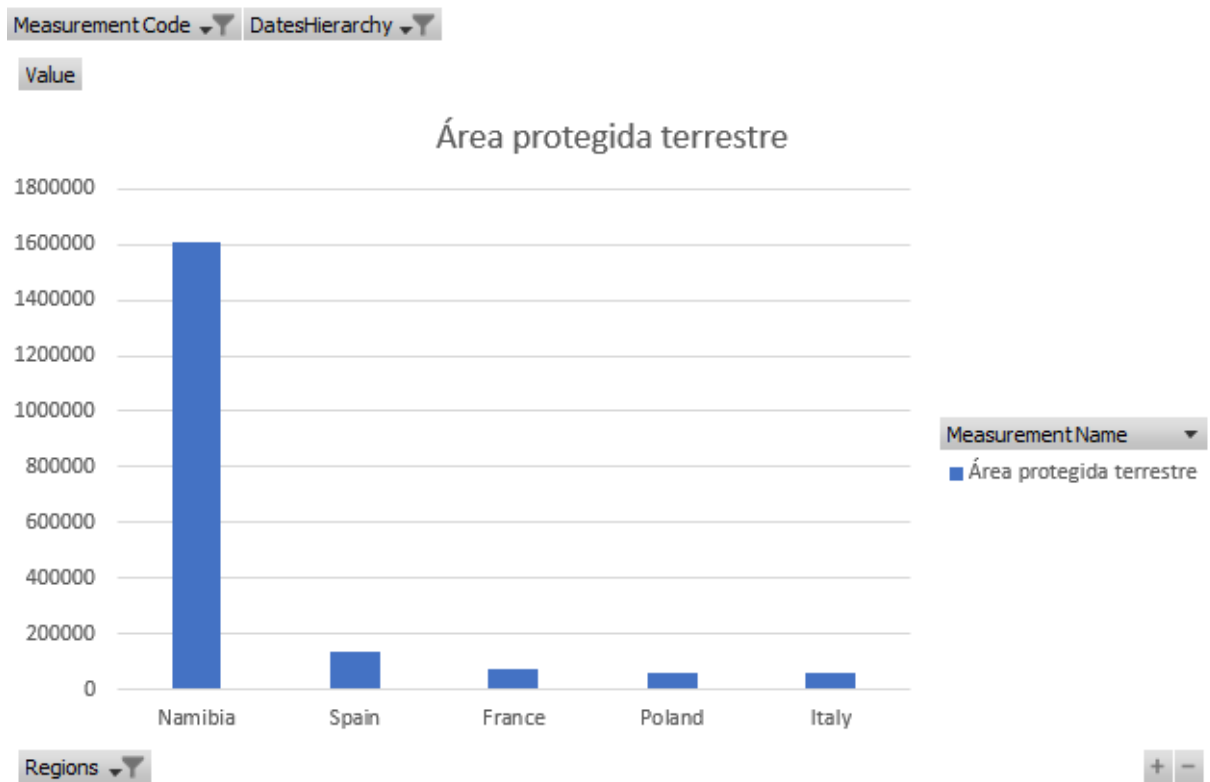


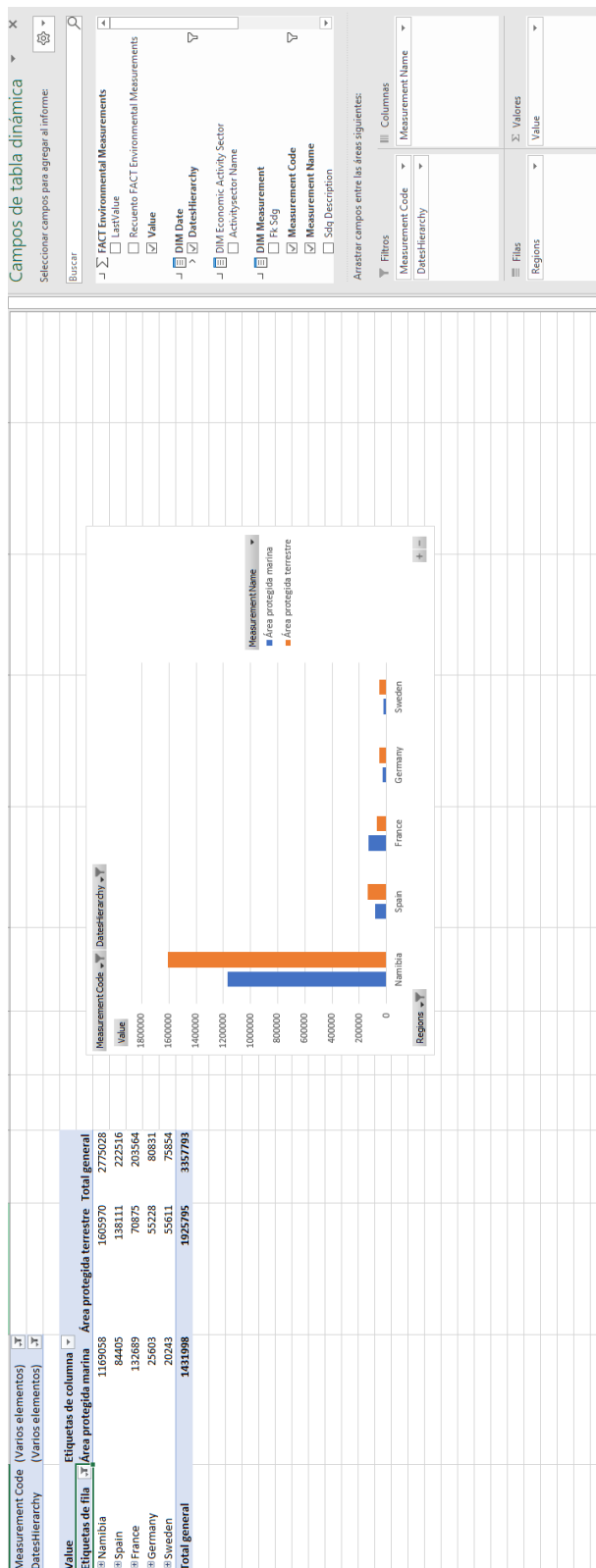
Destaca la aparición de Namibia como líder indiscutible del podio de países con mayor área protegida, tanto marina como terrestre. También resulta interesante observar cómo el resto de países que configuran el podio en las 3 gráficas son todos de Europa.

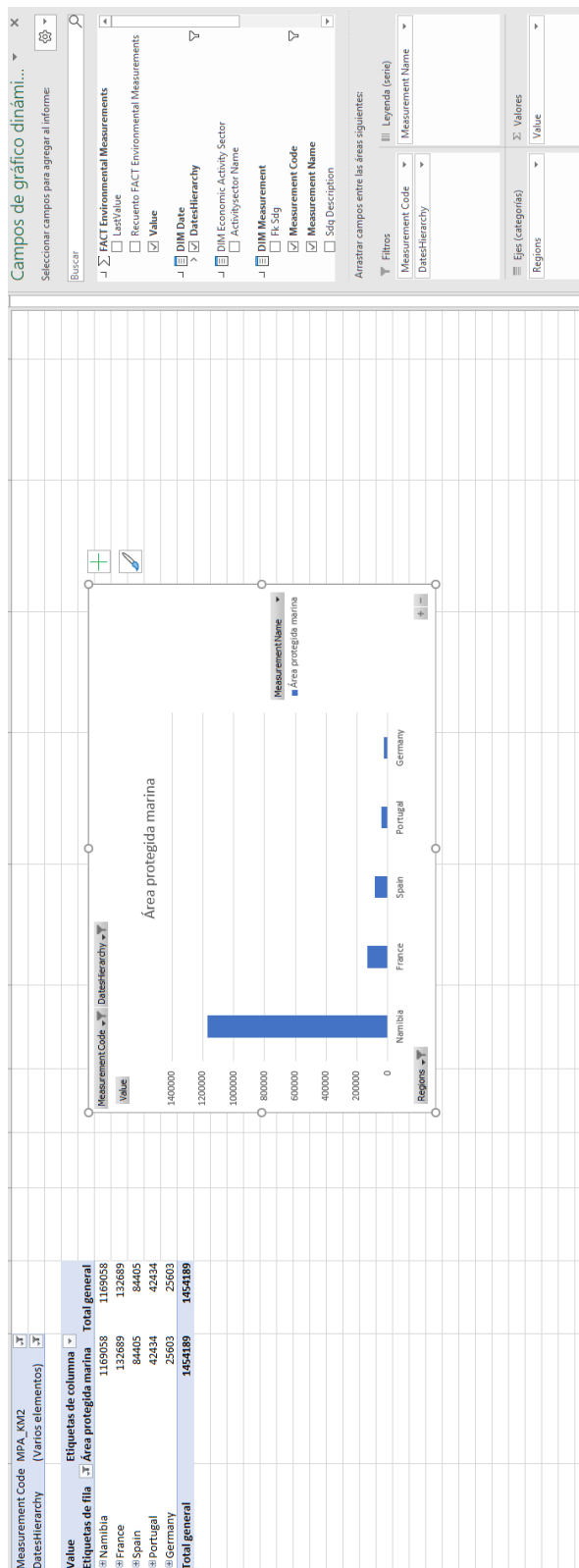
España y Francia aparecen en las 3 gráficas; Francia por delante de España en áreas marinas protegidas y España por delante de Francia en áreas terrestres protegidas.

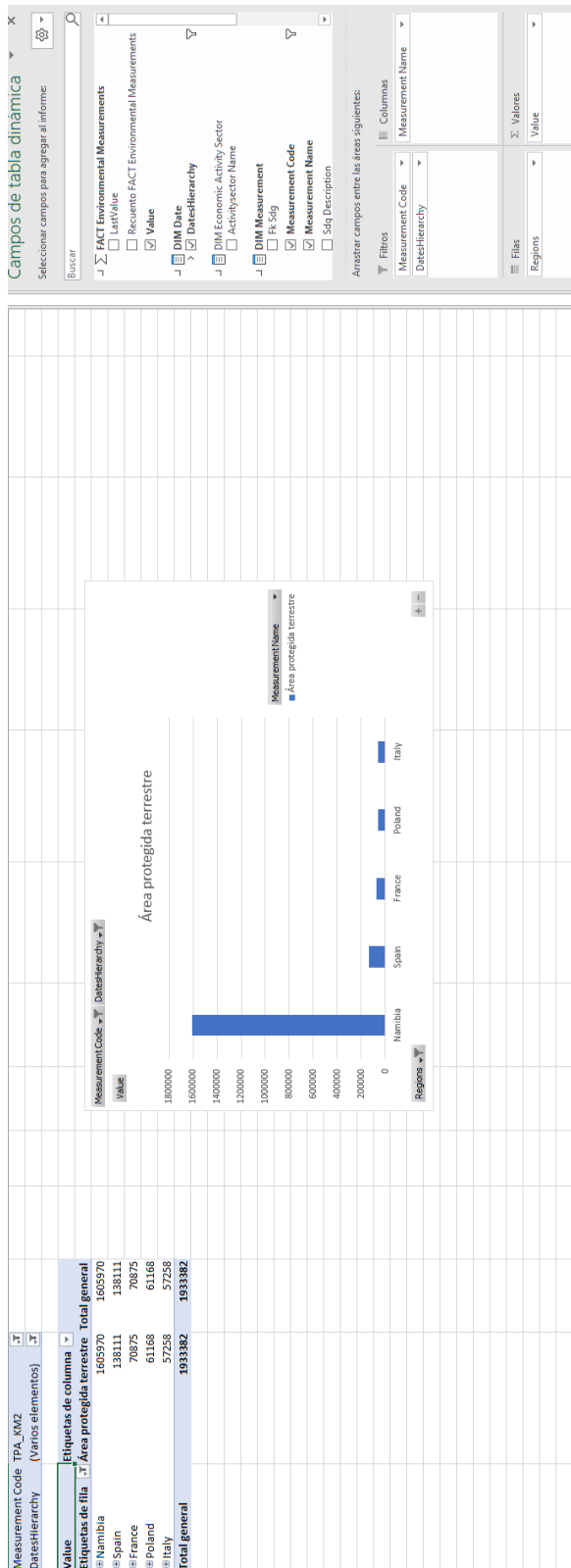


Cabe destacar también la inclusión de Suecia en el podio total, a pesar de no haber aparecido en ningún podio de áreas protegidas parciales, y la aparición de Polonia en cuarta posición en áreas terrestres protegidas. Se cae del podio total por no tener costa y por tanto, no poder aportar nada en áreas marinas protegidas.







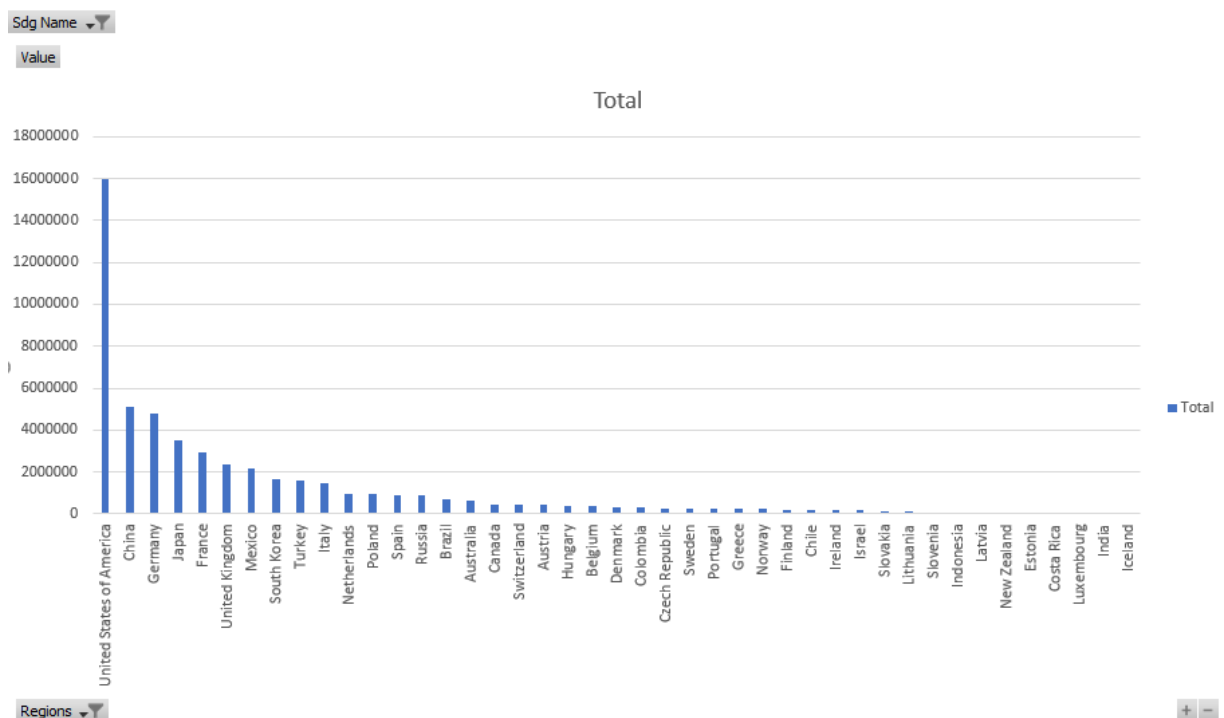


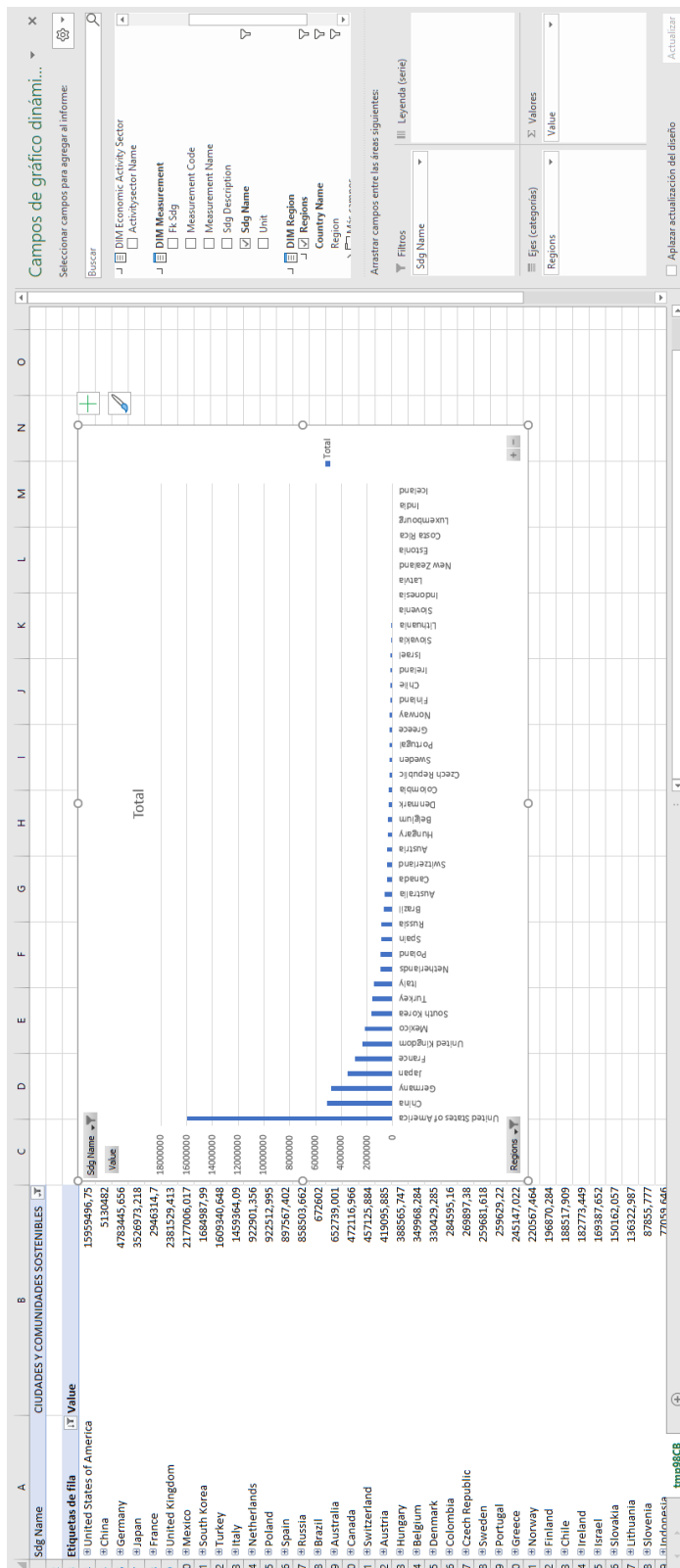
3.4. Identificación de los países con mayor impacto sobre el ODS «Ciudades y Comunidades sostenibles»

La consulta se ha realizado mediante Excel, utilizando el cubo de balances energéticos. Se ha fijado el objetivo de desarrollo sostenible (*SDG name*) y se han incluido los nombres de los países en el eje X. Por último, se han ordenado las etiquetas de fila por orden descendente según el valor del agregado.

Se observa que Estados Unidos y China, 2 de los países industrializados con mayor población y menos restricciones de polución, se encuentran en los primeros puestos.

También es interesante que Estados Unidos se desmarque del resto de países por un margen tan amplio, lo que puede significar que no se tienen tantos datos de otros países o Estados Unidos tenga datos incorrectos.

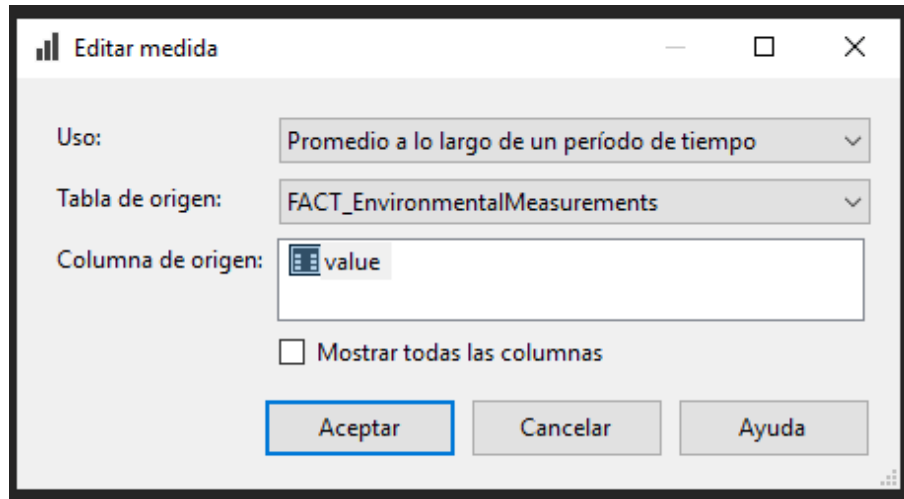


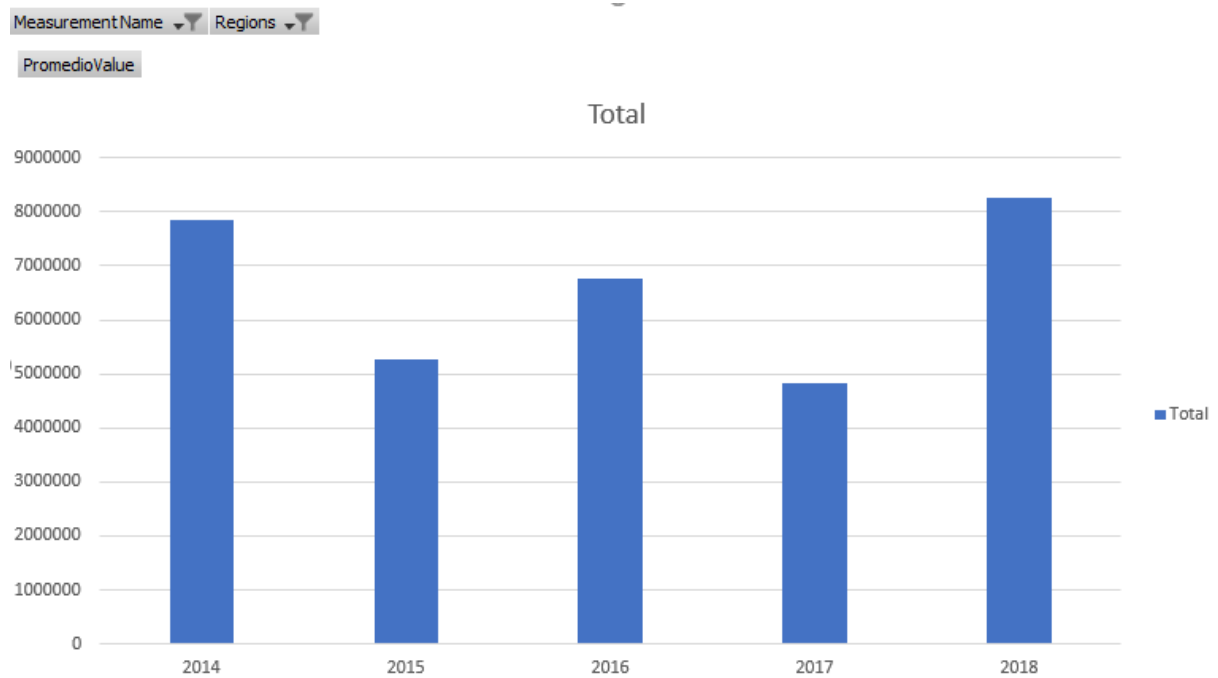


3.5. Análisis del promedio de inversión en el ámbito medioambiental de la gestión de aguas residuales de la Comunidad Valenciana (2014-2018)

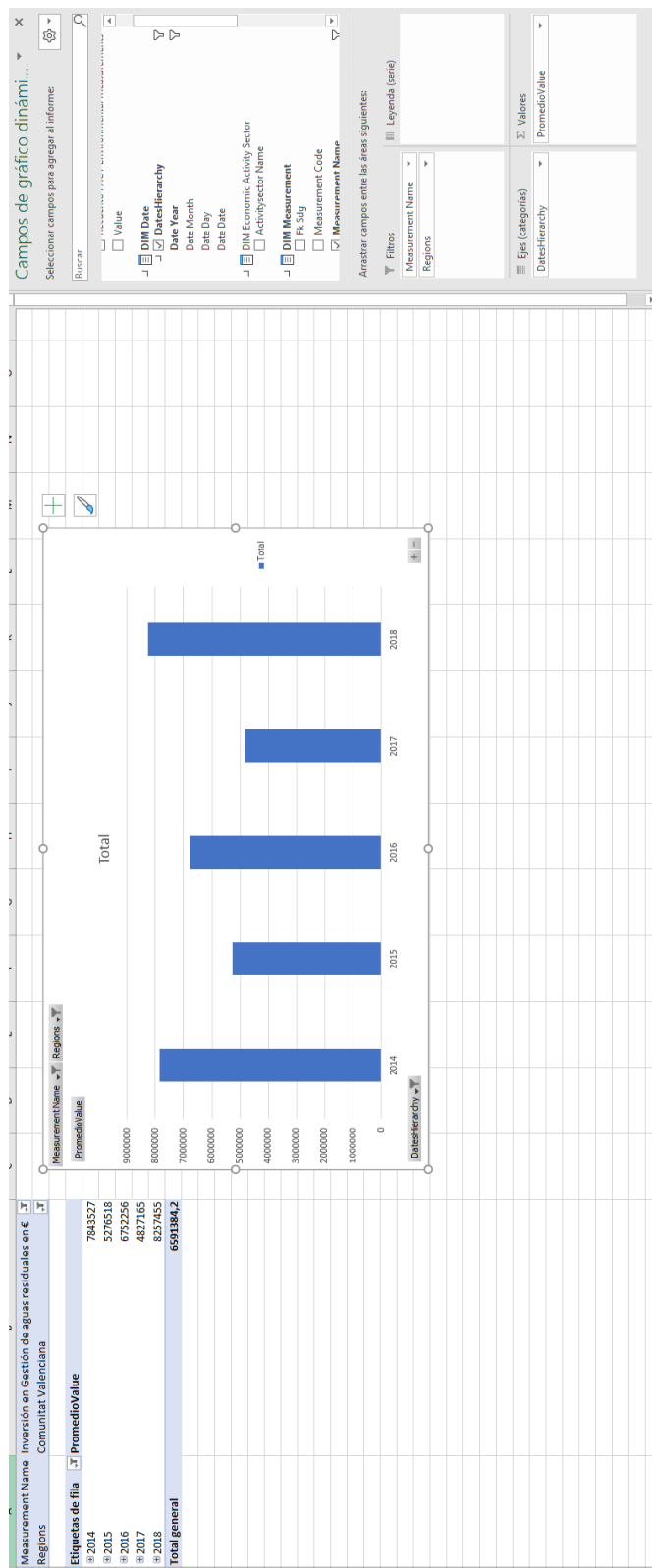
La consulta se ha realizado mediante Excel, utilizando el cubo de medidas ambientales. Se ha fijado la medida a analizar (*measurement name*) y la región, así como los años en que se desea analizar.

Para obtener el promedio se ha creado una nueva medida agregada en el cubo de medidas ambientales como promedio a lo largo de un período de tiempo.



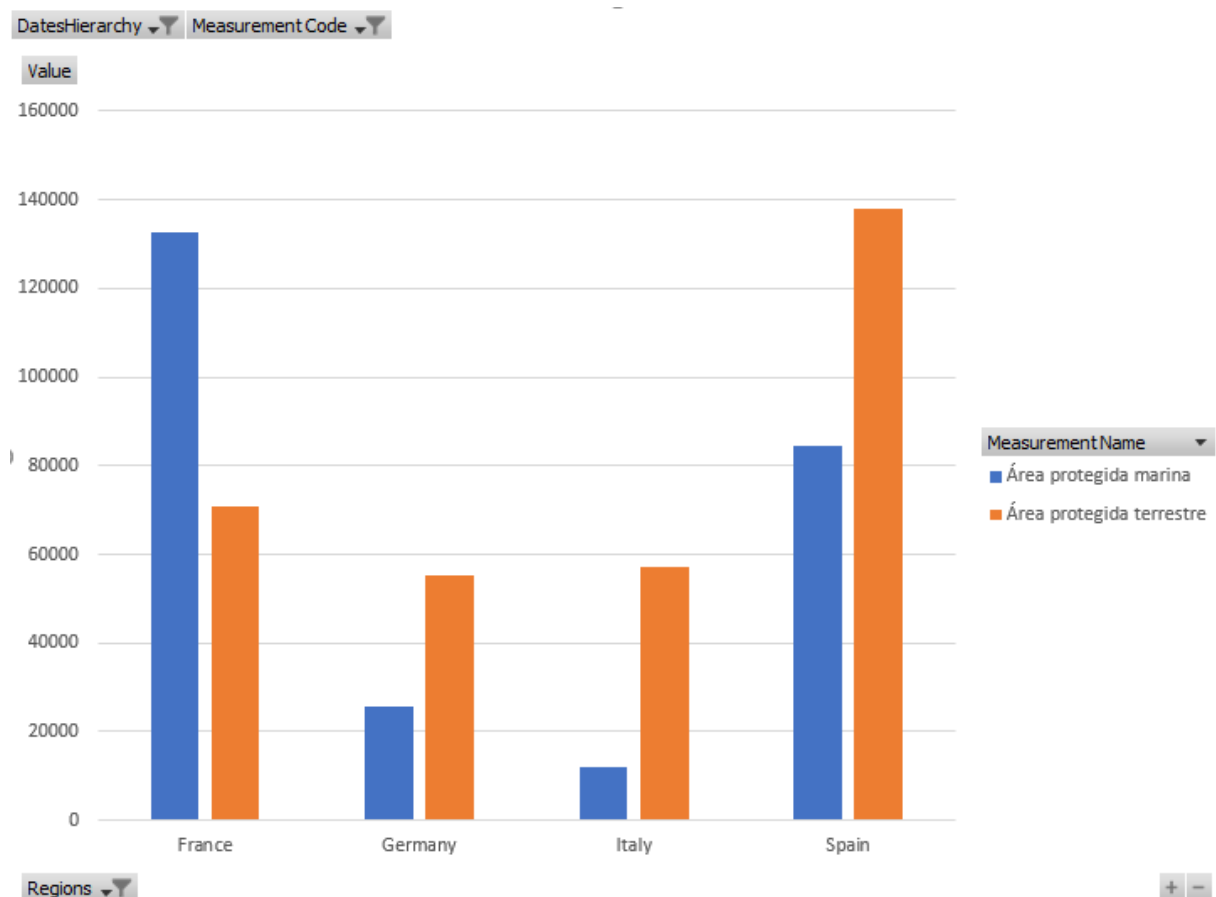


Se observa que la inversión fluctúa en torno a 6.000.000€, siendo 2018 el año con mayor inversión en esta medida ambiental.



3.6. Comparativa del área protegida entre Alemania, Francia, España e Italia (2019)

La consulta se ha realizado en Excel, utilizando el cubo de medidas ambientales. Se ha filtrado por regiones, seleccionando solo las de interés para este análisis, y se han añadido al eje del gráfico dinámico. Se han seleccionado a su vez los códigos de las medidas correspondientes a las áreas protegidas terrestre y marina.



A primera vista, choca el hecho de que Italia sea el país con menos área marina protegida, incluso menos que Alemania, el país con menos costa de los analizados. También es interesante observar la comparativa entre Francia y España.

Francia tiene más área protegida marina que España y España tiene más área protegida terrestre que Francia. Sin embargo, España tiene más porcentaje de frontera acuática que Francia, y Francia tiene una extensión terrestre mayor que España.

