Diseño y Construcción de Data Warehouse
Curso 2017 – Proyecto
Solución

Análisis Multidimensional de Datos de la Red de monitoreo de la calidad del aire en Montevideo

1 Requerimiento 1

Se quiere analizar información acerca de la polución en el aire a través del tiempo. Las mediciones se desean analizar según las líneas de ómnibus, la ubicación geográfica, los métodos de monitoreo y los contaminantes en el aire.

Interesa visualizar: la cantidad de líneas de ómnibus que circulan en cada barrio, la polución y la cantidad de contaminantes en el aire en dichos barrios. Esta información debe poder agruparse por fecha, mes y año.

Las líneas de ómnibus se identifican con un número (por ejemplo línea 117). De los métodos de monitoreo interesa su nombre y una descripción del mismo, por ejemplo el método UYMVD_O3 mide el ozono (O3) con un sensor electroquímico. Por otro lado, sobre los contaminantes en el aire interesa su nombre, y se quiere poder clasificarlos según su subtipo y su tipo. Por ejemplo, el subtipo del dióxido de azufre (SO2) es compuesto de azufre (S) y éste corresponde al tipo contaminante primario. Finalmente, la ubicación geográfica se refiere a la estación donde se realiza el monitoreo del aire, la cual se quiere clasificar según el barrio en el que se encuentra.

2 Requerimiento 2

Se quiere analizar información acerca de los hogares afectados por la polución a lo largo del año 2013, teniendo en cuenta su ubicación geográfica y los contaminantes en el aire.

Los hogares se desean clasificar según el tipo de vivienda, la forma de tenencia, el nivel de confort, problemas de la vivienda y de acuerdo a su ubicación (si está o no en un asentamiento).

Interesa visualizar la cantidad de hogares afectados, así como también la cantidad de personas, cantidad de personas menores de 14 años, cantidad de personas mayores de 14 años, cantidad de hombres y cantidad de mujeres. Estas cantidades se quieren ver tanto sumadas como promediadas al agrupar según los distintos criterios.

Los indicadores antes mencionados se quieren visualizar a lo largo del año 2013 y se deben poder totalizar por fecha, día de la semana, mes, trimestre y semestre. Además, según la estación del año (otoño, invierno, primavera y verano).

Los tipos de viviendas son casa, apartamento y otros. Las formas de tenencia de vivienda son propietario, inquilino y ocupante. Los tipos de problemas de las viviendas pueden ser muros agrietados, goteras en el techo, etc.

Los elementos de confort son si cuenta con lavavajilla, secadora de ropa, microondas, etc. El nivel de confort de la vivienda se debe deducir a partir de la cantidad de electrodomésticos que tenga el hogar. Se debe clasificar en 4 niveles de confort, definiendo un rango de cantidad de electrodomésticos para cada nivel. También se puede tener en cuenta alguna otra información, como si tiene servicio doméstico o no, o el origen del agua que utiliza.

La ubicación geográfica se desea agrupar según el barrio y el centro comunal.

3 Diseño Conceptual

En esta sección se presentan todos los pasos que conforman al diseño Conceptual.

3.1 Dimensiones y Medidas

Dimensiones

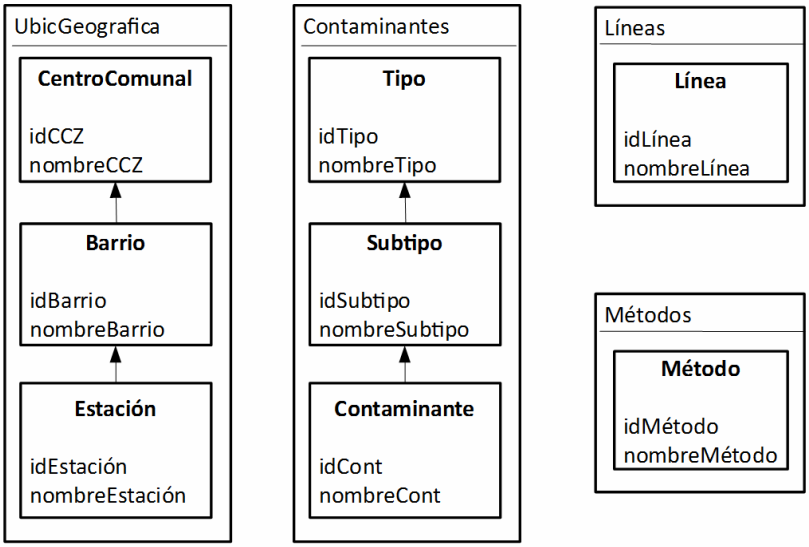

Figura 1.- Dimensiones *UbicGeografica*, *Contaminantes* y *Líneas*.

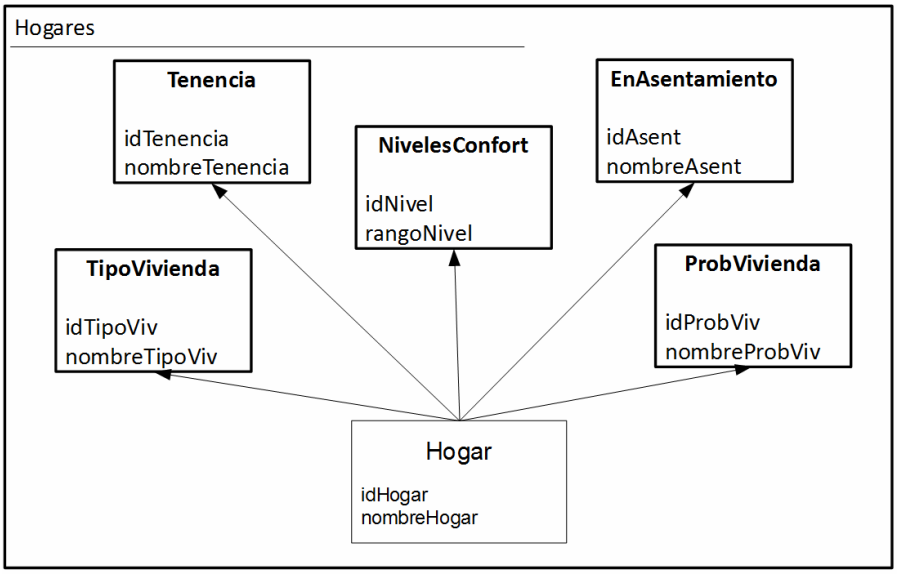

Figura 2.- Dimensión *Hogares*.

Para el caso de los niveles de confort se definen los siguientes rangos:

- entre 0 y 3 electrodomesticos nivel de confort 1
- entre 4 y 7 electrodomesticos nivel de confort 2
- entre 8 y 11 electrodomesticos nivel de confort 3
- entre 12 y 15 electrodomesticos nivel de confort 4
- más de 15 electrodomesticos nivel de confort 5
- más de 15 electrodomesticos y servicio doméstico nivel de confort 6

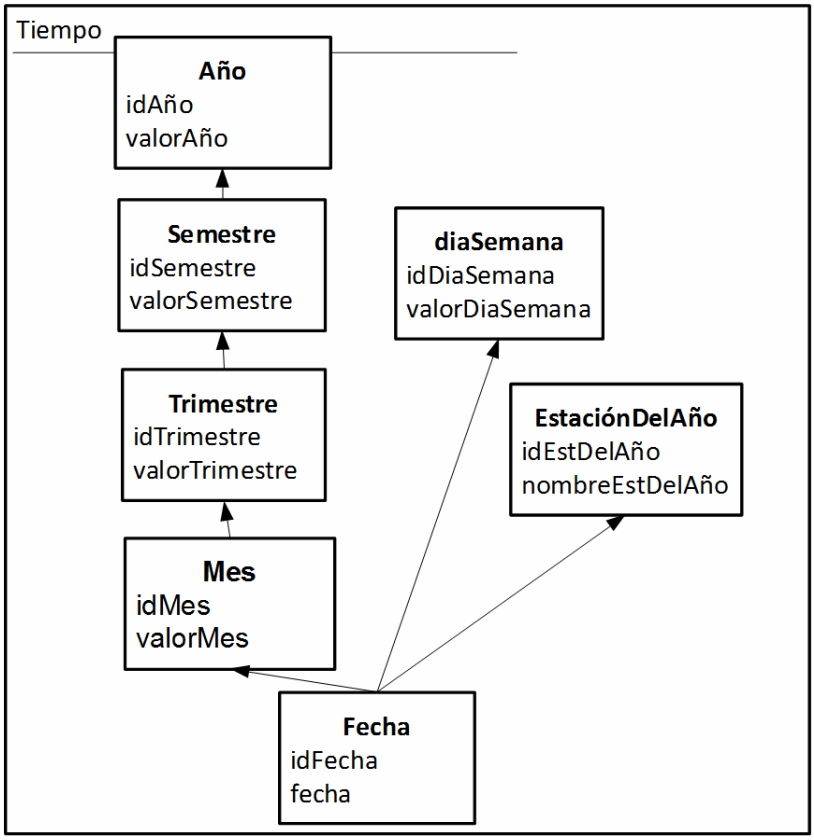

Figura 3.- Dimensión *Tiempo*.

Medidas

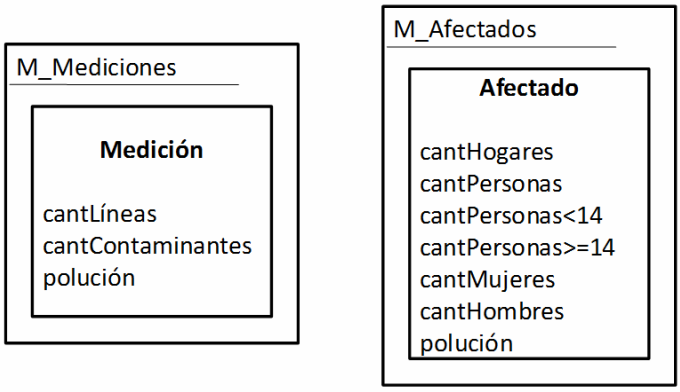

Figura 4.- Medidas *M_Mediciones* y *M_Afectados*.

La medida polución, si bien no está dentro de los requerimientos, es una medida interesante, por lo que se considera una medida opcional.

3.2 Tabla de requerimientos, dimensiones y medidas

A continuación se muestra el estudio de as dimensiones y medidas que forman parte de cada requerimiento. Una "X" en la tabla indica que la dimensión o la medida es necesaria para el requerimiento.

	Requerimiento 1	**Requerimiento 2**
Dimensiones		
UbicGeográfica	X	X
Tiempo	X	X
Líneas	X	
Métodos	X	
Contaminantes	X	X
Hogares		X
Medidas		
polución	X	X
cantLíneas	X	
cantContaminantes	X	
cantHogares		X
cantPersonas		X
cantPersonas>=14		X
cantPersonas<14		X
cantMujeres		X
cantHombres		X

Tabla 1.- Tabla de Requerimientos.

3.3 Relaciones dimensionales

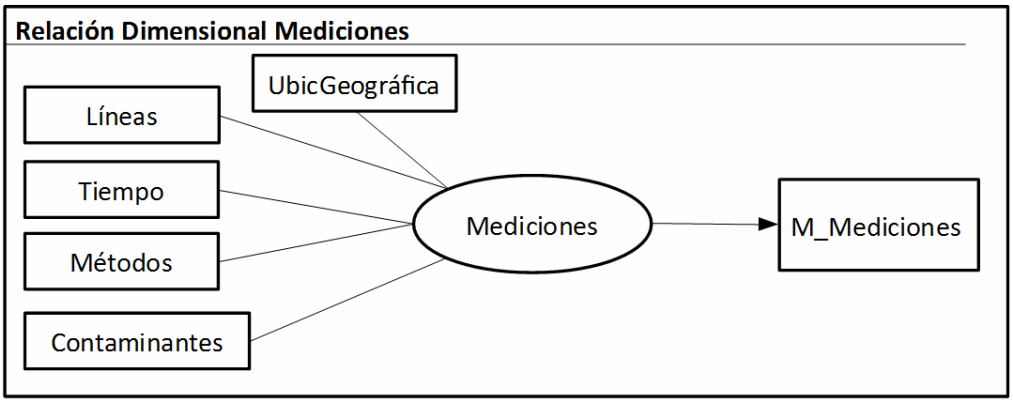

Figura 5.- Relación Dimensional *Mediciones*.

La medida *cantLíneas* solo interesa cuando las dimensiones *Métodos* y *Contaminantes* están en el nivel *All*.

Las medidas *cantContaminantes* y *polución* solo interesan cuando la dimensión *Líneas* está en el nivel *All*.

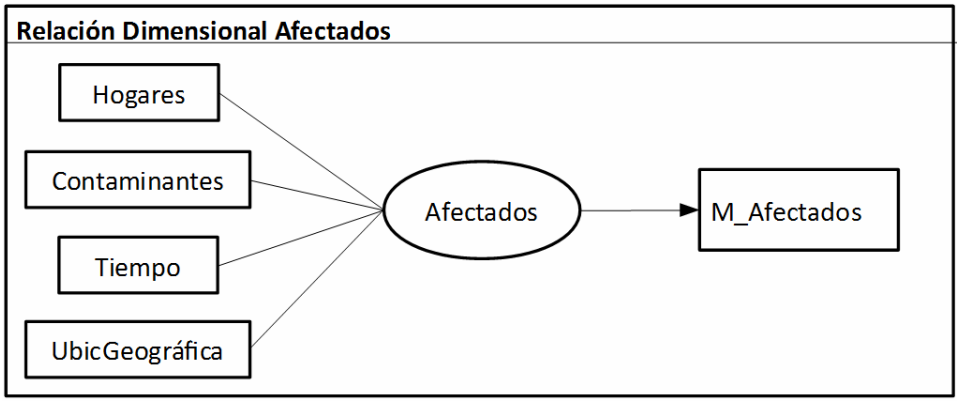

Figura 6.- Relación Dimensional *Afectados*.

La dimensión *UbicGeográfica* es considerada desde el nivel *Barrio* hacia arriba.

La medida *polución*, si bien no está dentro de los requerimientos, es una medida interesante, por lo que se considera una medida opcional.

3.4 Tablas de Roll-Up

Mediciones

		cantLineas	cantContaminantes	polución
Ubic.Geográfica	Estación --> Barrio	NoA, Prom	NoA, Prom	Prom
	Barrio --> CentroComunal	NoA, Prom	NoA, Prom	Prom
	CentroComunal --> All	NoA, Prom	NoA, Prom	Prom
Tiempo	Fecha --> Mes	NoA, Prom	NoA, Prom	Prom
	Mes --> Trimestre	NoA, Prom	NoA, Prom	Prom
	Trimestre --> Semestre	NoA, Prom	NoA, Prom	Prom
	Semestre --> Año	NoA, Prom	NoA, Prom	Prom
	Año --> All	NoA, Prom	NoA, Prom	Prom
	Fecha --> diaSemana	NoA, Prom	NoA, Prom	Prom
	diaSemana --> All	NoA, Prom	NoA, Prom	Prom
	Fecha --> EstaciónDelAño	NoA, Prom	NoA, Prom	Prom
	EstaciónDelAño --> All	NoA, Prom	NoA, Prom	Prom
Líneas	Línea --> All	A, Prom		
Método	Método --> All		NoA	Prom
Contaminantes	Contaminante --> Subtipo		A, Prom	Prom
	Subtipo --> Tipo		A, Prom	Prom
	Tipo --> All		A, Prom	Prom

Tabla 2.- Tabla de Roll-Up para la Relación Dimensional *Mediciones*.

Para el caso de la polución no tiene sentido considerar la aditividad porque es un índice, sin embargo, sí tiene sentido hallar el promedio de dichos valores.

Afectados

		cantHogares	cantPersonas	cantPersonas<14	cantPersonas>=14	cantMujeres	cantHombres
Ubic.Geográfica	Estación --> Barrio						
	Barrio --> CentroComunal	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	CentroComunal --> All	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
Tiempo	Fecha --> Mes	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Mes --> Trimestre	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Trimestre --> Semestre	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Semestre --> Año	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Año --> All	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Fecha --> diaSemana	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	diaSemana --> All	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Fecha --> EstaciónDelAño	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	EstaciónDelAño --> All	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
Contaminantes	Contaminante --> Subtipo	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Subtipo --> Tipo	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Tipo --> All	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
Hogares	Hogar --> TipoVivienda	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	TipoVivienda --> All	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	Hogar --> Tenencia	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	Tenencia --> All	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	Hogar --> EnAsentamiento	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	EnAsentamiento --> All	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	Hogar --> ProblemasViv	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	ProblemasViv --> All	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom	NoA, Prom
	Hogar --> Niveles	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom
	Niveles --> All	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom	A, Prom

Tabla 3.- Tabla de Roll-Up para la Relación Dimensional *Afectados*.

4 Diseño Lógico

A continuación se presentan las tablas definidas para la base de datos del DW.

Los datos de cada fecha son su identificador idFecha, el valor de dicha fecha, el identificador de la estación del año a la cual pertenece la fecha, el nombre de la estación, el identificador del día de la semana a la cual corresponde la fecha, su nombre, el número de mes correspondiente y su nombre, el trimestre, el semestre y el anio correspondiente. Todos estos datos se encuentran en el siguiente esquema relación:
tiempo(idFecha, fecha, idEstacionDelAnio, nomEstacionDelAnio, idDiaSemana, nomDiaSemana, idMes, mes, trimestre, semestre, anio)

Los datos de cada estación de monitoreo son su identificador idEstacion, su nombre, el identificador del barrio al cual pertenece, el nombre de dicho barrio, el identificador del centro comunal al cual pertenece el barrio de la estación y el nombre del centro comunal; y los mismos se encuentran en el siguiente esquema relación:
ubicacionGeog_1(idEstacion, nomEstacion, idBarrio, nomBarrio, idCCZ, nomCCZ)

Los datos de cada barrio son su identificador idBarrio, su nombre, el identificador del centro comunal al cual pertenece el barrio y el nombre del centro comunal; estos datos se encuentran en el siguiente esquema relación:
ubicacionGeog_2(idBarrio, nomBarrio, idCCZ, nomCCZ)

Los datos de cada línea de ómnibus son su identificador idLinea y su nombre; y los mismos se encuentran en el siguiente esquema relación:
lineas(idLinea, nomLinea)

Los datos de cada método aplicado para el monitoreo del aire son su identificador idMetodo y su nombre; y los mismos se encuentran en el siguiente esquema relación:
metodos(idMetodo, nomMetodo)

Los datos de cada contaminante son su identificador idContaminante, su nombre, el identificador del subtipo al cual pertenece dicho contaminante, el nombre del subtipo y el identificador del tipo de contaminante con el nombre del tipo correspondiente. Estos datos se encuentran en el siguiente esquema relación:
contaminantes(idContaminante, nomContaminante, idSubTipo, nomSubTipo, idTipo, nomTipo)

Los datos de cada hogar son su identificador idhogar, su nombre (asignado por la encuestadora), el identificador del nivel de confort asociado, el nombre del nivel de confort, el identificador y valor asociado a su pertenencia o no a un asentamiento, el identificador del tipo de tenencia del hogar, el nombre del tipo de tenencia, el identificador del tipo de vivienda y el nombre asociado a dicho tipo. Estos datos se encuentran en el siguiente esquema relación:
hogares(idHogar, nomHogar, idNivelConfort, nomNivelConfort, idAsentamiento, nomAsentamiento, idTenencia, nomTenencia, idTipoVivienda, nomTipoVivienda)

Los datos de los tipos de problemas que pueden presentar los hogares son su identificador idProblema y su nombre; los mismos se encuentran en el siguiente esquema relación:
problemasEnHogares(idProblema, nomProblema)

la relación entre los niveles tal y cual de una de las jerarquías de Hogares, es N a N, se propone esta solución para el esquema relacional.

Dado que cada hogar podría presentar varios problemas, la relación entre los niveles *Hogar* y *ProbVivienda* de una de las jerarquías de la dimensión *Hogares* (figura 2), es N a N. Por lo tanto, como solución a este problema, se propone crear una *bridge table* para el esquema relacional. De esta forma, es posible relacionar cada hogar con cada uno de los problemas que presenta. Estos datos se encuentran en el siguiente esquema relación:
bridgeHogaresProblemas(idHogar, idProblema)

En el siguiente esquema relación se presenta información de cada uno de las mediciones realizadas:
mediciones(idEstacion, idMetodo, idContaminante, idLinea, idFecha, polución)

En el siguiente esquema relación se presenta información de cada uno de los hogares afectados:
afectados(idHogar, idBarrio, idContaminante, idFecha, cantPersonas, cantPersonas>=14, cantPersonas<14, cantMujeres, cantHombres, polución)

En esta base de datos se cumplen las siguientes restricciones de inclusión:

$\pi_{idHogar}$(bridgeHogaresProblemas) $\subseteq$ $\pi_{idHogar}$(hogares)
$\pi_{idProblema}$(bridgeHogaresProblemas) $\subseteq$ $\pi_{idProblema}$(problemasEnHogares)

$\pi_{idEstacion}$(mediciones) $\subseteq$ $\pi_{idEstacion}$(ubicacionGeog_1)
$\pi_{idMetodo}$(mediciones) $\subseteq$ $\pi_{idMetodo}$(metodos)
$\pi_{idContaminante}$(mediciones) $\subseteq$ $\pi_{idContaminante}$(contaminantes)
$\pi_{idLinea}$(mediciones) $\subseteq$ $\pi_{idLinea}$(lineas)
$\pi_{idFecha}$(mediciones) $\subseteq$ $\pi_{idFecha}$(tiempo)

$\pi_{idhogar}$(afectados) $\subseteq$ $\pi_{idHogar}$(hogares)
$\pi_{idBarrio}$(afectados) $\subseteq$ $\pi_{idEstacion}$(ubicacionGeog_2)
$\pi_{idContaminante}$(afectados) $\subseteq$ $\pi_{idContaminante}$(contaminantes)
$\pi_{idFecha}$(afectados) $\subseteq$ $\pi_{idFecha}$(tiempo)

En la figura 7 y en la figura 8 se muestra la representación gráfica de las relaciones de la base de datos.

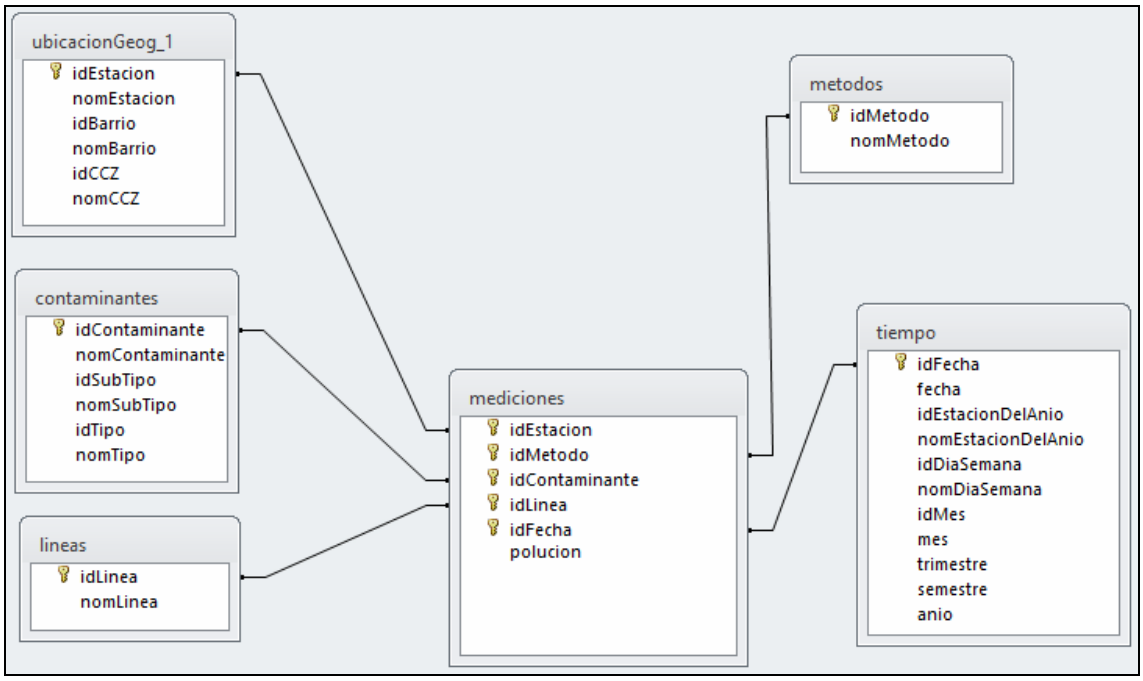

Figura 7.- Representación de la relación correspondiente a la relación dimensional *Mediciones*.

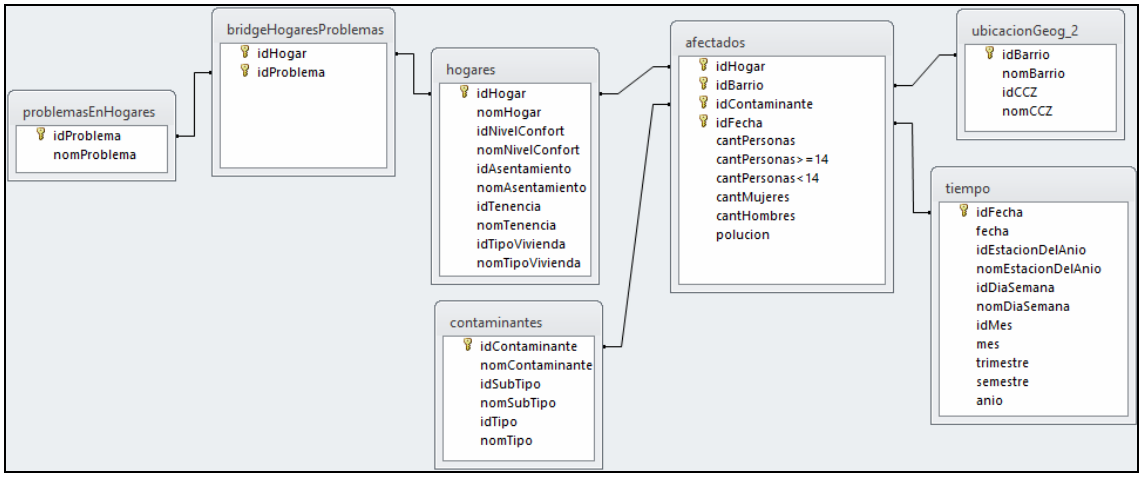

Figura 8.- Representación de la relación correspondiente a la relación dimensional *Afectados*.