VII Reunión sobre casos prácticos de inspección y vigilancia de mercados y entidades.

Santiago de Chile

DATA MINING
CONCEPTOS Y EXPERIENCIA EN LA FISCALIZACIÓN DEL MERCADO DE VALORES DE CHILE

Marcelo García R

Sonia Muñoz C.

Santiago, 17 de mayo de 2011

DATA MINING. CONCEPTOS

La Inteligencia de negocios o **Business Intelligence** (BI) se puede definir como un proceso de análisis de datos acumulados con el objeto **de extraer una cierta inteligencia o conocimiento de ellos.**

En la categoría de datos acumulados se incluyen las distintas bases de datos de que administran las empresas, las cuales contienen información de sus clientes, cadenas de suministros, ventas, etc.

En el ámbito de la **fiscalización del mercado de valores**, las bases de datos corresponden a:

- Registros de transacciones
- Registros de órdenes de compraventa
- Registros de custodia de acciones
- Registros de directores, etc.

DATA MINING. CONCEPTOS

La tecnología **Business Intelligence** no es nueva. Ha estado presente de varias formas por lo menos en los últimos **20 años**, comenzando por generadores de reportes y sistemas de información ejecutiva en los 80's.

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de los computadores, como a su bajo costo de almacenamiento.

Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de **información oculta**, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información.

DATA MINING. CONCEPTOS

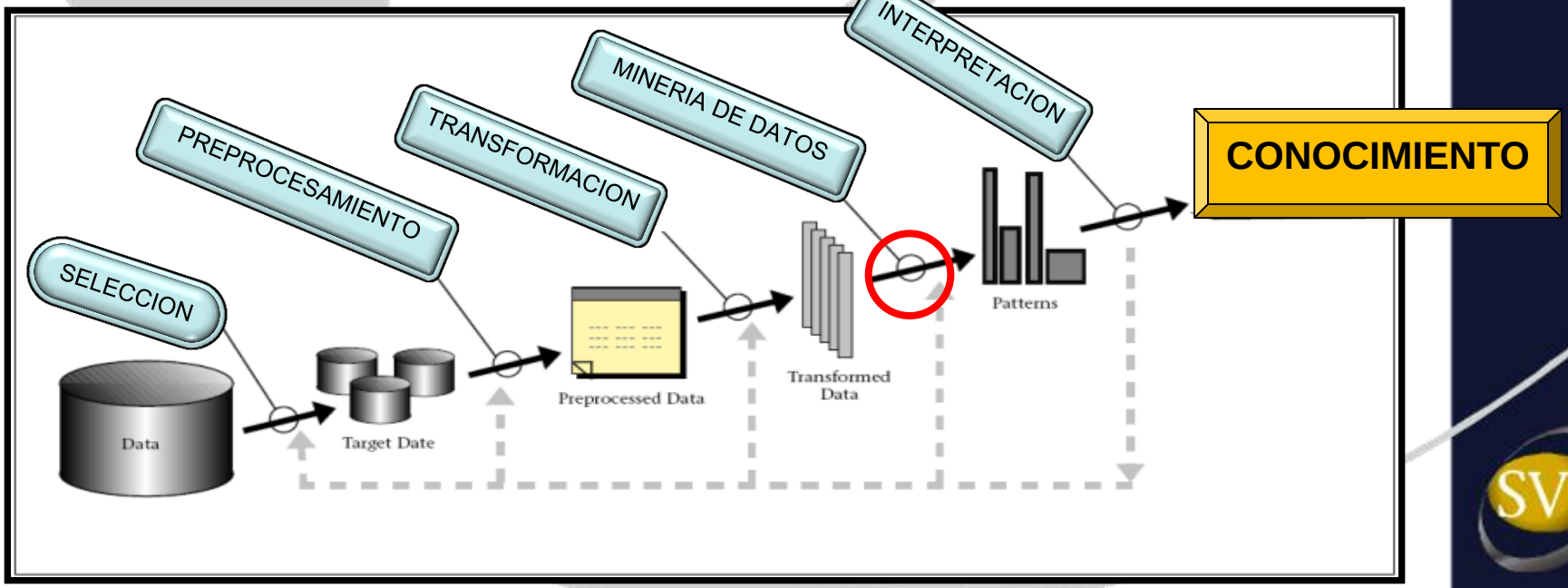
El descubrimiento de esta información oculta es posible gracias a la **Minería de Datos** (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad

Pero es el **descubrimiento del conocimiento en bases de datos** (KDD, Knowledge Discovery from Databases) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

DATA MINING. CONCEPTOS

De forma general, los **datos son la materia prima bruta**. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información.

Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento.



Data Mining. Conceptos

La capacidad de generar y almacenar información creció considerablemente en los últimos tiempos, se ha estimado que la cantidad de datos en el mundo almacenados en bases de **datos se duplica cada 20 meses**. Es así que hoy las organizaciones tienen gran cantidad de datos almacenados y organizados, pero a los cuales no les pueden analizar eficientemente en su totalidad.

Con las sentencias SQL se puede realizar un **primer análisis**, aproximadamente el **80%** de la información se obtiene con estas técnicas. El 20% restante, que la mayoría de las veces, contiene la información más importante, requiere la utilización de técnicas más avanzadas.

DATA MINING. CONCEPTOS

El **Descubrimiento de Conocimiento en Bases de Datos** (KDD) apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos, de esta manera permitirá al usuario el uso de esta información valiosa para su conveniencia.

El KDD es el Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos .

El objetivo del KDD es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos. Otro aspecto es que la interacción humano-máquina deberá ser flexible, dinámica y colaboradora.

El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los datos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos.

DATA MINING. CONCEPTOS

El proceso de **descubrimiento de conocimiento en bases de datos** involucra varios pasos:

1. **Determinar las fuentes de información** que pueden ser útiles y dónde conseguirlas.
2. **Diseñar** el esquema de almacenamiento de datos (Data Warehouse)
3. **Implantación** del almacenamiento de datos que permita la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
4. **Selección, limpieza y transformación** de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).
5. **Seleccionar y aplicar** el método de minería de datos apropiado. Esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc.
6. **Evaluación, interpretación, transformación y representación** de los resultados extraídos.
7. **Difusión y uso** del nuevo conocimiento. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

Data Mining. Conceptos

Las metas del KDD son:

- **Procesar automáticamente** grandes cantidades de datos crudos.
- **Identificar los patrones** más significativos y relevantes.
- Presentarlos como **conocimiento** apropiado para satisfacer las metas del usuario.

IMPORTANCIA DE LA INFORMACIÓN EN LA FISCALIZACIÓN DEL MERCADO DE VALORES.

Bajo el concepto de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación (retail).

IMPORTANCIA DE LA INFORMACIÓN EN LA FISCALIZACIÓN DEL MERCADO DE VALORES.

Un proceso típico de **minería de datos** consta de los siguientes pasos generales:

- **Selección del conjunto de datos**
- **Análisis de las propiedades de los datos**

  Realizar una validación de los datos, observando presencia de valores atípicos, ausencia de datos (valores nulos), etc.
- **Transformación del conjunto de datos de entrada**, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como **preprocesamiento** de los datos.
- **Seleccionar y aplicar la técnica de minería de datos**, se construye el modelo predictivo, de clasificación o segmentación.
- **Extracción de conocimiento**, mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesado diferente de los datos.
- **Interpretación y evaluación de datos**, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

DATA MINING. APLICACIÓN

La SVS cuenta con el software **Modeler SPSS IBM 14.1 (ex Clementine)** con el cual han desarrollado aplicaciones que permiten procesar grandes cantidades de datos crudos.

1. Obtención de la **red familiar** directa de los inversionistas analizados.(*)
2. Acceso, **sin límite de registros,** a bases de transacciones y datos pre-validados.
3. Realización de **cruces de bases de datos** para la detección de uso de información privilegiada.
4. Modelo para calcular el “aporte individual de cada inversionista a la **formación del precio** de un instrumento y periodo determinado”.
5. Obtención instantánea de **estadísticas** asociadas a las bases de datos, que permiten determinar la habitualidad en los las operaciones.

DATA MINING. APLICACIÓN

Respecto de los beneficios y/o mejoras obtenidos con el software Modeler SPSS IBM 14.1 pueden destacarse los siguientes:

1. **Acceso directo a las bases de datos** de la SVS (según perfil de usuario), sin el requisito previo de tener conocimientos de programación.
2. **Manejo ilimitado de registros**. La base de datos de transacciones almacena decenas de millones de registros a los cuales se puede accesar.
3. **Seguridad en la manipulación de datos**, disminuyendo riesgos de pérdida y/o alteración de los datos almacenados.
4. **Requerimiento de menor tiempo** para procesar importantes volúmenes de información.
5. **Mayor conocimiento** de la información almacenada en la SVS.

DATA MINING. APLICACIÓN

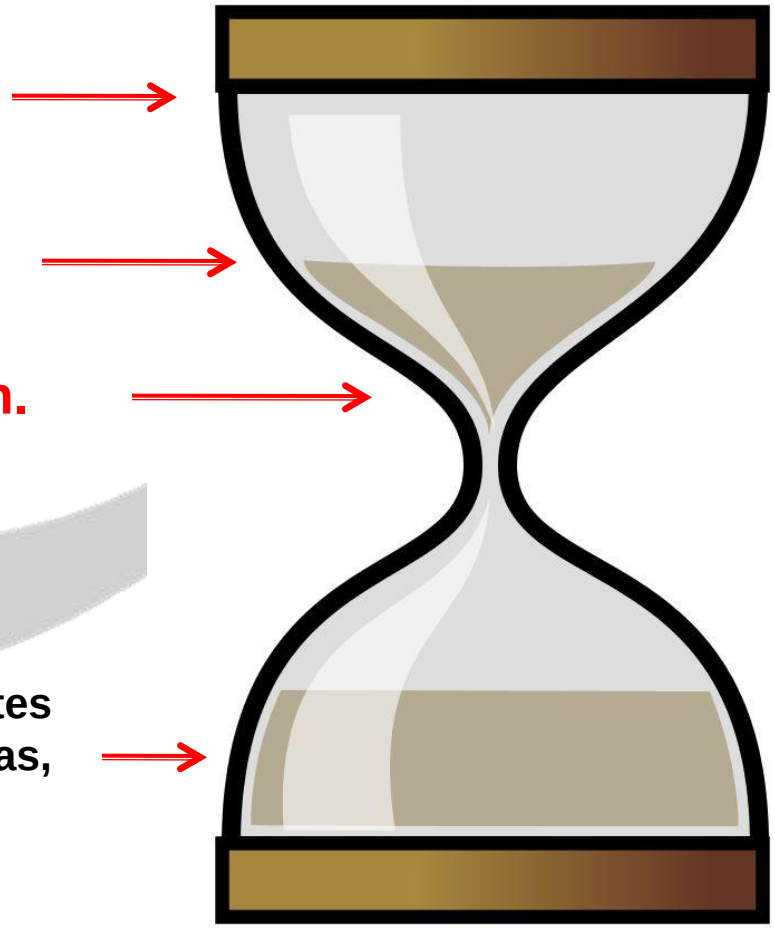
ESQUEMA DE MANEJO DE INFORMACIÓN

Requerimiento de transacciones (acciones, RF, etc)

Pre-procesamiento de los datos (determinación del periodo crítico, inversionistas sospechosos, etc)

Análisis de la información. Minería de datos

Nuevo requerimiento de antecedentes para respaldar operaciones (facturas, ficha de cliente)



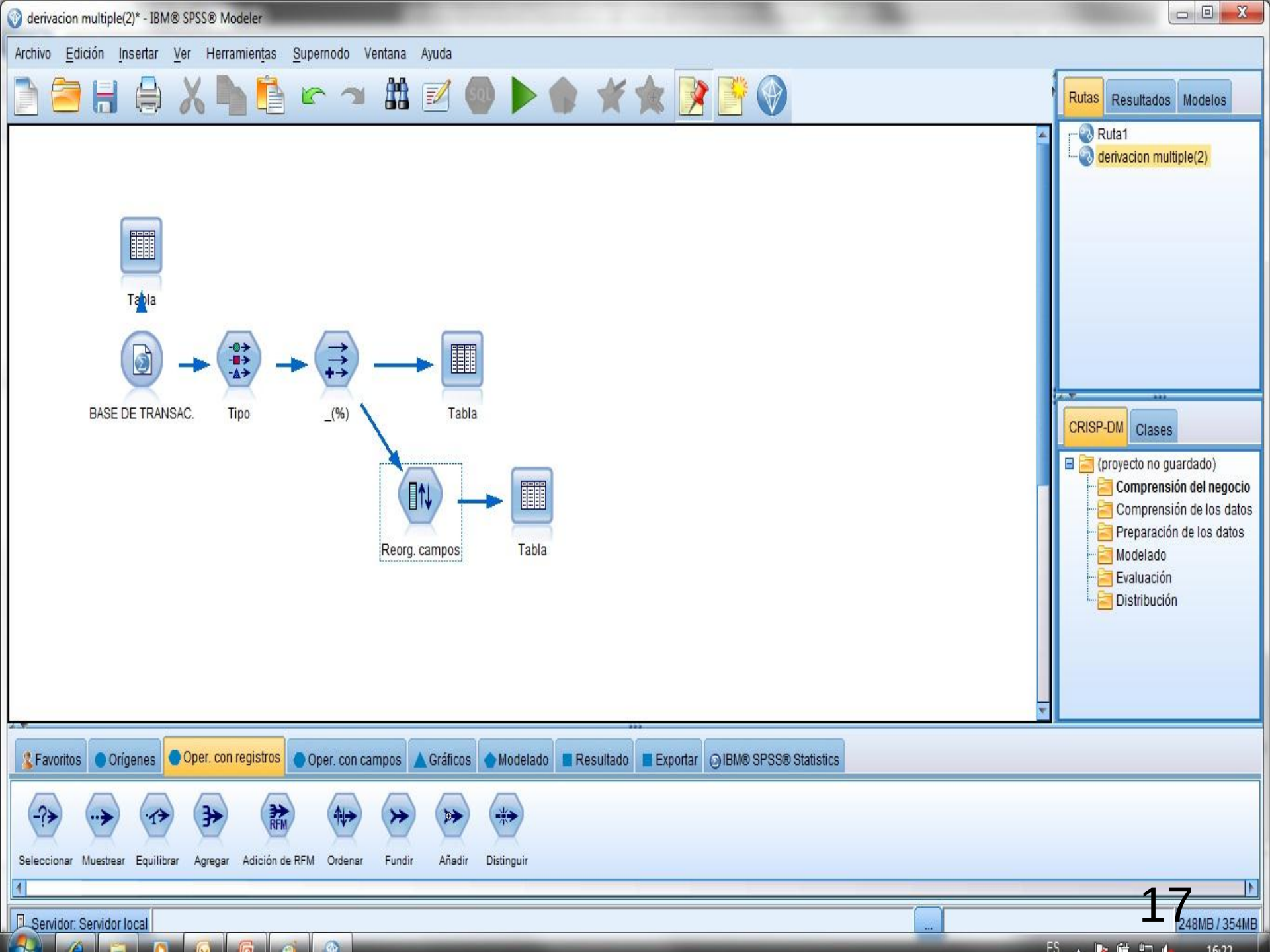
Modeler SPSS IBM 14.1

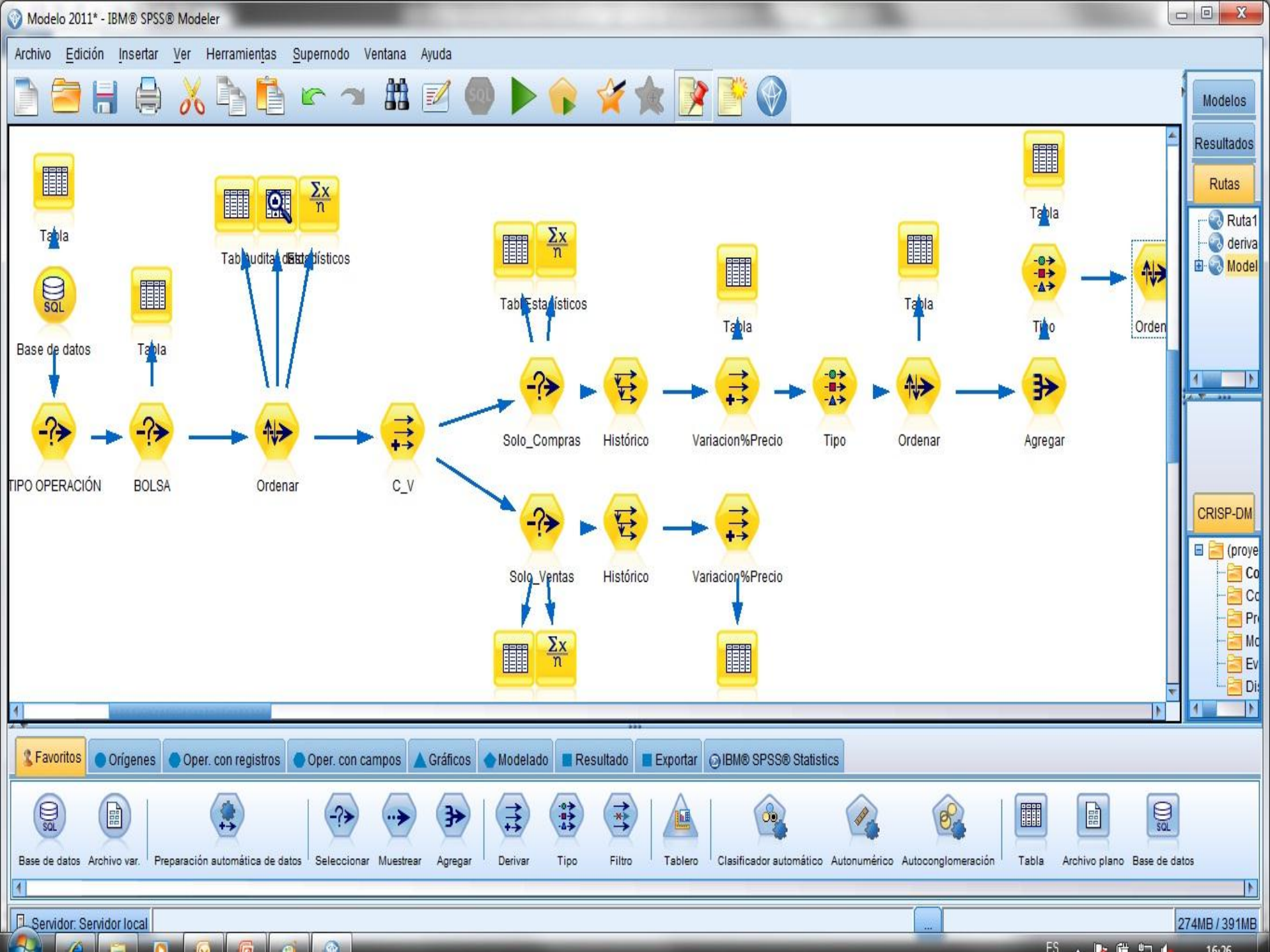
En relación a este tema, en el año 2005 se evaluó alternativas de software en el mercado, decidiendo la compra del software **Modeler SPSS IBM 14.1 (ex Clementine),** el cual renovamos en el año 2010.

COSTO PROYECTO SOFTWARE MINERIA DE DATOS	
	US $
2 licencias concurrentes	122.049,60
Capacitación	20.380,04
Servidor	6.491,45
TOTAL	**142.308,15**

Data Mining. Aplicación

Modeler SPSS IBM 14.1

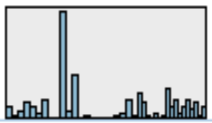
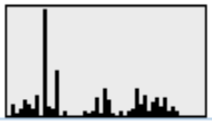
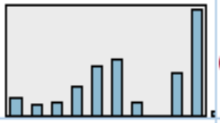
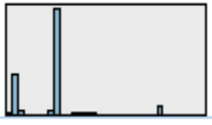
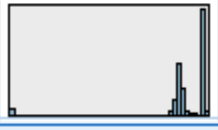

derivacion multiple(2)* - IBM® SPSS® Modeler
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
Rutas Resultados Modelos
Ruta1
derivacion multiple(2)
Tabla
BASE DE TRANSAC.
Tipo
_(%)
Tabla
Reorg. campos
Tabla
CRISP-DM Clases
(proyecto no guardado)
Comprensión del negocio
Comprensión de los datos
Preparación de los datos
Modelado
Evaluación
Distribución
Favoritos Orígenes Oper. con registros Oper. con campos Gráficos Modelado Resultado Exportar IBM® SPSS® Statistics
Seleccionar Muestrear Equilibrar Agregar Adición de RFM Ordenar Fundir Añadir Distinguir
Servidor: Servidor local
248MB / 354MB


Modelo 2011* - IBM® SPSS® Modeler
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
Modelos
Resultados
Rutas
Ruta1
deriva
Model
CRISP-DM
Tabla
Base de datos
Tabla
Tabla
Tabla
Tabla
Tabla
Tabla
Tipo
Orden
TIPO OPERACIÓN
BOLSA
Ordenar
C_V
Solo_Compras
Histórico
Variacion%Precio
Tipo
Ordenar
Agregar
Solo_Ventas
Histórico
Variacion%Precio
Favoritos
Orígenes
Oper. con registros
Oper. con campos
Gráficos
Modelado
Resultado
Exportar
IBM® SPSS® Statistics
Base de datos
Archivo var.
Preparación automática de datos
Seleccionar
Muestrear
Agregar
Derivar
Tipo
Filtro
Tablero
Clasificador automático
Autonumérico
Autoconglomeración
Tabla
Archivo plano
Base de datos
Servidor: Servidor local
274MB / 391MB

Auditar datos de [16 campos]

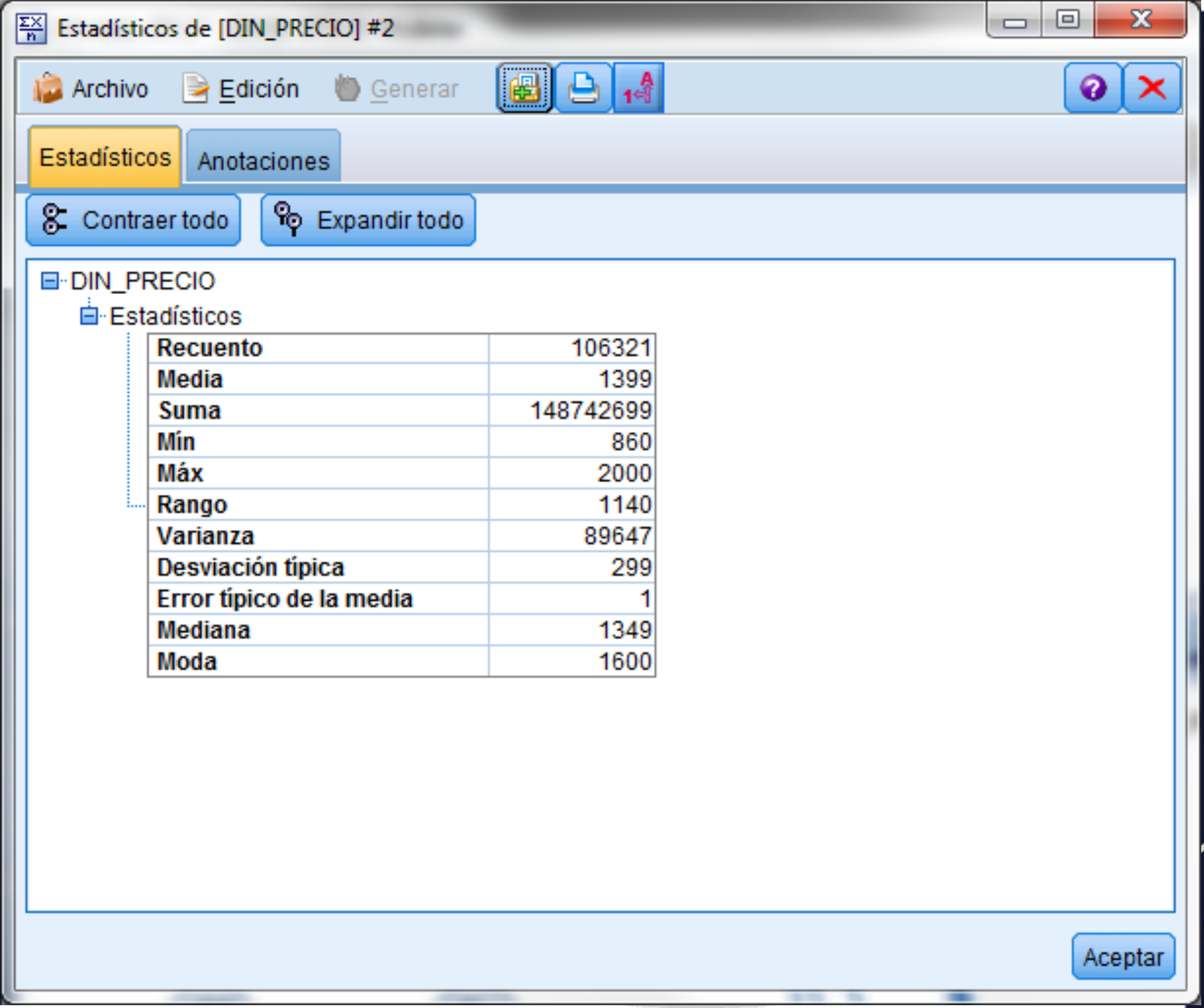
Archivo Edición Generar

Auditar | Calidad | Anotaciones

Campo	Gráfico de muest...	Medida	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
DIN_OFI_N...		Continua	3333	3376	3353	13	0	--	263541
DIN_RUTIN...		Nominal	76306360	99555580	--	--	--	43	263541
DIN_DIGIN...		Nominal	--	--	--	--	--	11	263541
DIN_FECHA		Continua	20100104	20110110	20101626	2672	3	--	263541
DIN_HORA		Continua	0	999999	127026	24496	-0	--	263541
DIN_NUMB...		Continua	1	3	1	1	3	--	263541
DIN_FOLIO		Continua	1	999999	195983	138641	2	--	263541
DIN_RUTIN...		Continua	0	99555580	88196801	13940008	-4	--	263541

[1] Indica un resultado de varios modos [2] Indica un resultado muestreado

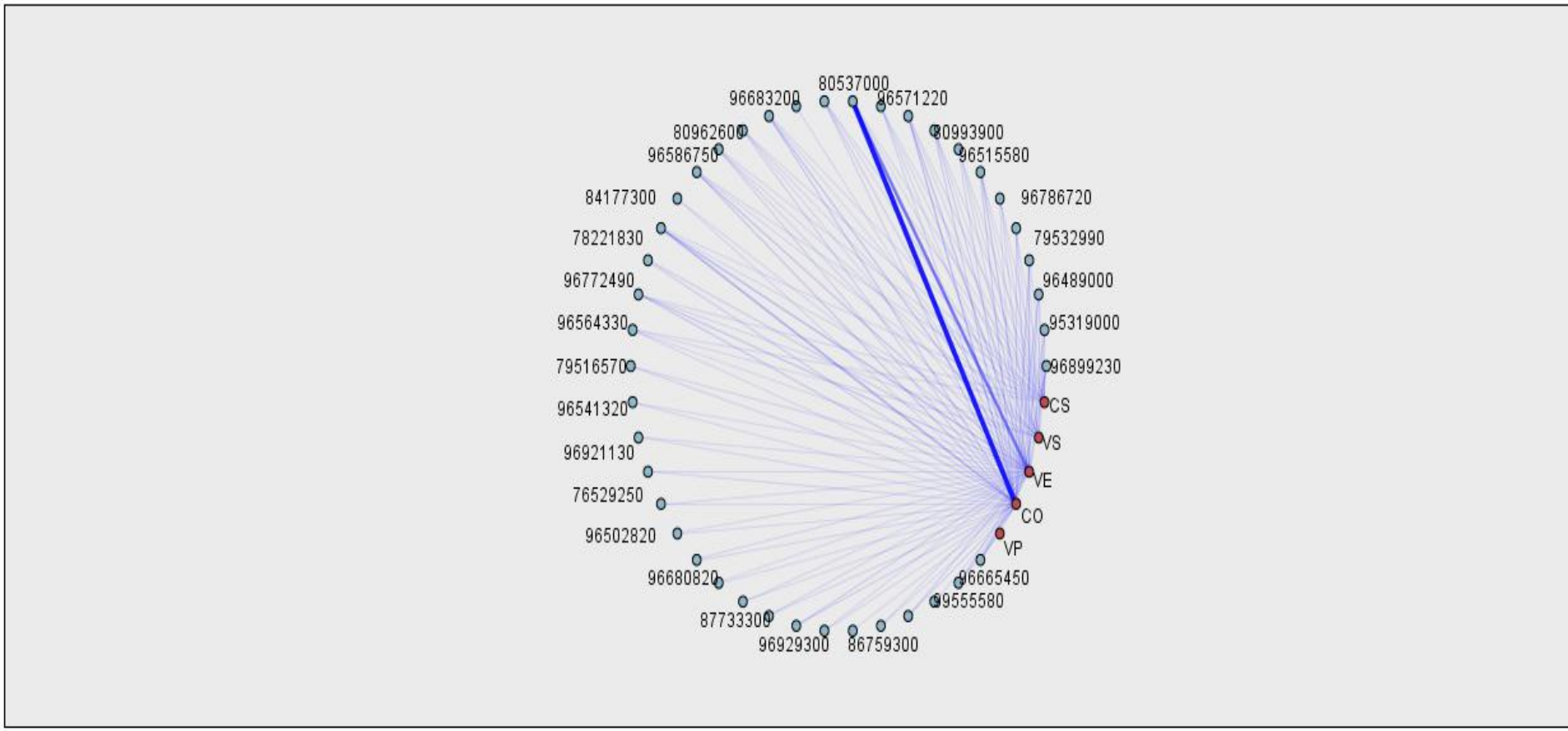
Aceptar

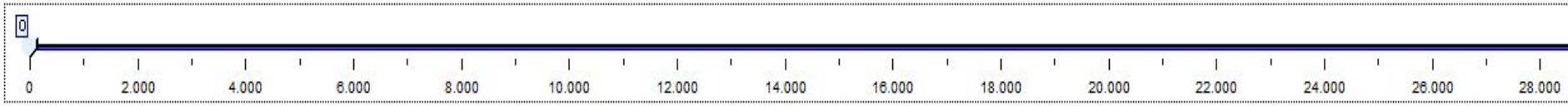

Estadísticos de [DIN_PRECIO] #2
Archivo
Edición
Generar
Estadísticos
Anotaciones
Contraer todo
Expandir todo
DIN_PRECIO
Estadísticos
Recuento 106321
Media 1399
Suma 148742699
Mín 860
Máx 2000
Rango 1140
Varianza 89647
Desviación típica 299
Error típico de la media 1
Mediana 1349
Moda 1600
Aceptar

Malla de DIN_RUTINTERM x DIN_TIPOPERC : Absolutos

Archivo Edición Generar Ver Malla Enlace

Gráfico Anotaciones





VII Reunión sobre casos prácticos de inspección y vigilancia de mercados y entidades.

Santiago de Chile

DATA MINING
CONCEPTOS Y EXPERIENCIA EN LA FISCALIZACIÓN DEL MERCADO DE VALORES DE CHILE

Marcelo García R

Sonia Muñoz C.

Santiago, 17 de mayo de 2011