

# **Departamento Administrativo Nacional de Estadística**

Dirección de Metodología y Producción Estadística



## **METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES**

JUNIO 2002

---

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR- EXPO-MET-01 VERSIÓN : 01 Página 1 DE 28 Fecha: 24-06-02
Elaboro: Estadísticos consultores	Reviso: Secretario Técnico	Aprobó: Director DIMPE

## INFORME DE AVANCE\*

### Resumen<sup>1</sup>

Uno de los mayores problemas que afrontamos los técnicos en el desarrollo de la investigación en general, y particularmente en la económica, es la baja confiabilidad que se tiene sobre los datos, lo cual es el resultado de mala calidad en algunos o todos los procesos de acopio, grabación y crítica de la información. A pesar de ser rigurosos en la utilización de las herramientas estadísticas en los procesos de análisis económico muchos de nuestros modelos están afectados claramente por la baja calidad de la información utilizada para su elaboración. En este sentido, popularmente se dice que a un modelo que se alimenta con basura, éste de nuevo arroja basura. Enmarcado el DANE en esta preocupación, y como un esfuerzo mas dentro de los que viene realizando de tiempo atrás, se ha dado a la tarea de establecer mecanismos técnicos que permitan controlar la calidad de la información de comercio exterior, particularmente la correspondiente a las exportaciones. Para ello ha diseñado un nuevo sistema de captura de la información, el cual va a permitir automatizar muchos de los procesos manuales de control de calidad de la información de exportaciones, que están actualmente vigentes. En esta línea de trabajo, se ha diseñado un sistema de control de calidad para el valor por unidad de kilo de mercancía (valor unitario), el cual parte de controlar la validación de códigos de las posiciones arancelarias, validación de las unidades de medida para cada una de éstas, lo mismo que el número máximo de kilos admisibles para las posiciones que se declaran en unidades, terminando con la validación del valor unitario en dólares para cada posición. El presente documento describe la metodología desarrollada para tal fin.

\* Este estudio se desarrolla a partir del análisis de los registros de exportación correspondientes a 1996, haciendo parte integral o insumo del nuevo sistema de captura de la información de exportaciones, que actualmente desarrolla la Dirección de Estadísticas Básicas del DANE.

<sup>1</sup> Los autores agradecen la colaboración prestada por Edgar Amézquita, Pablo Mahecha y Mireya Rendón, en el levantamiento de la información y la conformación de los archivos. Las opiniones presentadas aquí no reflejan necesariamente las del DANE, siendo éstas entera responsabilidad de los autores.

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :2 DE 28 Fecha: 24-06-02
---	--	--

## INDICE

Introducción	5
Objetivos	6
I. MARCO TEORICO	6
I.1 Detección de datos atípicos o outliers	7
I.1.1 Qué es un outliers?	8
I.1.2 Efectos de los outliers	8
I.1.3 Influencia de los outliers en las estimaciones de los estadísticos	9
I.1.4 Análisis de Regresión	9
I.1.5 Análisis de residuales	10
II. METODOLOGIA	12
II.1 Validación de la codificación de las posiciones arancelarias	14
II.2 Validación de los códigos de unidad	14
II.3 Desemglobe del peso en kilos brutos o del valor Fob en dolares	15
II.3.1 Metodología para estimación de kilos	16
II.3.2 Metodología para estimación de fob dólares	17
II.4 Valor de número máximo de kilos para posiciones arancelarias con unidad de medida: número de unidades	19
II.5 Rango del valor unitario en dólares	20
II.5.1 Metodología de regresión	21
II.5.2 Metodología de los residuales estudentizados	23
II.5.3 Comparación de las dos metodologías	27
II.5.4 Período considerado en el estudio	28
II.5.5 Información necesaria para la implementación de las metodologías	28
II.5.6 Actualización de los archivos maestros	29
III. RESULTADOS	29
IV. CONCLUSIONES Y RECOMENDACIONES	30
BIBLIOGRAFIA	30

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :3 DE 28 Fecha: 24-06-02
---	--	--

## INTRODUCCION

Por suerte para la humanidad la toma de decisiones ha venido cambiando en su forma y estilo. Ahora es poco lo que se le deja a la mera intuición, o lo que denominaría Keynes, a los "espíritus animales" de los individuos. Los sistemas de información cubren todo tipo de elemento informático. Conbinan las diferentes formas de expresión de la información: datos, video y sonido. Utilizan la última tecnología existente para su transmisión: cable óptico. Y en forma general desbordan las barreras culturales, ideológicas, políticas y sociales, vía internet. En medio de tanto desarrollo, de tanta ciencia, de tanto invento, de nuevo volvemos al principio: al dato; elemento básico de la información; elemento reductor de la expresión de situaciones; elemento medidor del tamaño de los fenómenos en estudio; y principalmente elemento de información.

Esta reflexión nos lleva a pensar sobre un tema bastante trillado pero poco o nada manejado, que es el de la calidad del dato. Cuanto de lo que nos informan los datos no es real? A que distancia nos encontramos entre lo que informa el dato y lo real? Estas y muchas más preguntas nos podemos hacer respecto al problema de la calidad de los datos, siendo, tal vez, muchas de ellas de difícil respuesta, e incluso, de no respuesta, lo cual nos impulsa a tratar con seriedad esta problemática, pero a tratarla ya, iniciado con ello un camino largo pero productivo hacia el mejoramiento de las estadísticas.

Atraído por estos y otros tantos elementos técnicos, el DANE se da a la tarea de implementar un nuevo sistema de captura de la información de exportaciones, el cual en primera instancia actualiza el viejo y tal vez absoluto sistema de captura, transformando procesos hasta ahora manuales, en ágiles sistemas de control de calidad y consistencia de la información.

Para ello, se plantea la combinación entre controles normales de validación de códigos, tales como las posiciones arancelarias y las unidades de medida de cada una de ellas y el control de la información sobre el valor de la mercancía y el número de kilos, integrado y sintetizado en un rango admisible para el valor unitario, expresado en dólares, para cada una de las posiciones arancelarias, y por ende para cada uno de los registros de exportación en el momento de su grabación.

En este orden de ideas, el documento presenta la metodología elaborada para tal fin, inspirada en la identificación de valores atípicos o outliers. Teniendo en cuenta que la información a lo largo del año presenta grandes variaciones respecto a sus valores absolutos, y que muchas posiciones arancelarias tienen un número extremadamente reducido de registro en dicho período, se determino utilizar la información de exportaciones de todo el año de 1996 para diseñar la metodología y establecer los rangos de valores unitarios admisibles.

Por rangos se entenderá el intervalo generado por el valor máximo y mínimo admisibles para una determinada partida arancelaria. Bajo la claridad de que el valor mínimo es igual a cero. Esta metodología se implementara en un programa, utilizando las herramientas que brinda el paquete estadístico SAS, de tal forma que se pueda

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :4 DE 28 Fecha: 24-06-02
---	--	--

ejecutar periódicamente, por lo cual siempre se tendrá unos rangos actualizados y acordes a la evolución del comercio exterior del año anterior.

## OBJETIVO GENERAL

Elaborar una metodología tendiente al control de la calidad de la información de comercio exterior, basada en el análisis de los valores unitarios.

### Objetivos específicos

1. Conformar un archivo que contenga las posiciones arancelarias vigentes para el año de 1996, con los respectivos códigos de unidad de medida, según lo establecido para la legislación Colombiana.
2. Elaborar una metodología estadística para la determinación del número de kilos admisibles para las posiciones arancelarias cuya medida de unidad corresponde a número de unidades.
3. Elaborar una metodología estadística para la determinación de un rango de valores admisibles para los valores unitarios en dólares, de las partidas arancelarias del movimiento de comercio exterior de exportaciones.

## I.MARCO TEORICO

Antes de entrar a presentar los elementos teóricos y prácticos de los cuales se partió para el desarrollo de la metodología, es interesante discutir el problema que se está abordando.

Al hablar sobre la calidad de la información, en forma genérica, estamos entrando en un campo un tanto amplio y peligrosamente ambiguo, debido a la existencia de un espectro de enfoques que se le pueden dar al tema de la calidad. En este orden de ideas, la calidad de la información cubre todas las etapas de su formación, no siendo siempre posible tener un control sobre cada una de ellas, ya sea por razones netamente físicas, razones de costo, o razones inherentes al problema que estamos tratando. Es así, que nuestra intervención está limitada al control de una porción de las etapas de generación de la información, que seguramente van ha estar relacionadas con el control de los datos.

Ahora si bien es cierto que estamos limitando nuestro campo de acción, el problema se traslada a la identificación de los factores que intervienen en la calidad de los datos, factores que están asociados con unos actores que intervienen en la generación de los mismos. De nuevo debemos entrar a la identificación de los factores sobre los cuales tenemos posibilidad de intervención, que siendo justos y realistas, son definitivamente reducidos, pero determinantes en el logro de datos de buena calidad.

Con este ejercicio mental hemos reducido, un poco, la amplitud del problema a tratar. Nos enfrentamos ahora con una lista de factores sobre los cuales diferentes actores

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :5 DE 28 Fecha: 24-06-02
---	--	--

han intervenido a lo largo de la generación del dato, unos en forma indirecta o conceptual, y otros en forma directa u operacional. Este es precisamente el esquema de funciona en la generación de los datos de comercio exterior. Existen unos factores institucionales que nacen como respuesta a una legislación cambiante a través del tiempo, los cuales en la mayoría de los casos son de carácter fiscal, sin tener en cuenta las implicaciones estadísticas. Por otra parte existen otros factores no institucionales y de operacionalización de los procesos establecidos por los primeros, que nacen desde el mismo momento del diligenciamiento de los instrumentos del comercio exterior.

Pasados los procesos administrativos, en los cuales se engendran los datos, y sobre los cuales, y de la misma forma que con los factores institucionales, no intervienen los técnicos directamente<sup>2</sup> se entra en la etapa de crítica, grabación y procesamiento de información, factores sobre los cuales el tamaño de lo técnico toma su máxima relevancia.

Y hacia donde va este esquema?. Este esquema se dirige a resaltar las limitaciones con las cuales se enfrenta el trabajo técnico, ya que antes de tener su turno de intervención sobre el dato factores como el lavado de dólares del narcotráfico, la subfacturación o sobrefacturación, el contrabando técnico, la evasión de impuestos, los vicios administrativos, las deficiencias de los procesos administrativos preestablecidos, y otros factores, han tenido efectos nefastos sobre la calidad del dato, quedándole a la caja de herramientas mágicas de lo técnico, la tarea de depurar o establecer mecanismos que aseguren la calidad de los mismos.


Es aquí donde lo técnico se encuentra frágil ante la calamidad producida por factores exógenos a su resorte, siendo muy importante tener conciencia sobre los posibles alcances que tienen los procesos técnicos que buscan garantizar la calidad de los datos, que en procesos contiguos a la crítica, grabación y depuración de éstos, le dan vida a la información.

Tomando muy en cuenta los puntos anteriores, la intervención que tiene la presente metodología está dirigida a aportar una herramienta técnica que ayude, por un lado a lograr que la información sea trasladada, con total fidelidad, del documento de exportación a los archivos magnéticos del DANE, y por otro a la cualificación de la calidad de la información como tal.

Para facilitar la comprensión de la metodología, a continuación se presentan los elementos teóricos y prácticos de los cuales se partió, tales como los conceptos básicos de Análisis de Regresión, Análisis de Residuales y Detección de datos típicos.

## 1.1 Detección de datos atípicos o outliers

<sup>2</sup> Es posible que los técnicos intervengan en algunas etapas relacionadas con lo institucional, pero es sabido que sobre los conceptos técnicos están las decisiones políticas, no siendo siempre en la misma dirección. En la parte administrativa lo técnico se desvanece, y lo operativo toma su mayor dimensión.

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :6 DE 28 Fecha: 24-06-02
---	--	--

Establecidos los procesos de crítica de los datos, lo mismo que los de su grabación, los procesos sobre control de calidad de los mismos, de alguna manera se reducen a la identificación de datos atípicos o outliers, que sugieren sospecha sobre su veracidad, siendo este el enfoque que tiene la presente metodología.

El tema relacionado con outlier es antiguo, y sin lugar a dudas, ha sido una fuente de preocupación de los más estudiosos de la matemática y la estadística, que datan de dos siglos atrás. Comentarios realizados por Bernoulli indican que la practica de destacar observaciones discordantes es común desde hace alrededor de los 200años.

El mayor énfasis de esta revisión de outliers esta en análisis de regresión, como se discutirá en la metodología. Un outlier en pocas palabras es una observación extrema, cuyo residual al realizar un análisis de regresión supera en tres o cuatro desviaciones estándar la media. Dependiendo de su localización en el n-espacio los outlier pueden tener un moderado o severo efecto en el análisis de regresión.

#### I.1.1. Que es un Outlier.

Ninguna observación puede ser garantizada como una manifestación totalmente dependiente del fenómeno bajo estudio. Un evento de un caso en un millón puede suceder sin que esto nos pueda sorprender por las frecuencias asociadas a este evento. Intuitivamente, la realidad probable de una observación es reflejado con la relación de otras observaciones que fueron obtenidas bajo condiciones similares. Las observaciones que en la opinión del investigador se encuentran aparte del grupo de los datos son llamados "outliers", "observaciones discordantes", "valores contaminantes", etc.

Otras definiciones que se les han dado a los outliers a través del tiempo son:

"Es aquella observación que presenta diferencia con respecto a la frecuencia asociada de esta y de otras observaciones con las que son combinadas".

"Es una observación que aparece marcadamente desviada de los otros miembros de la muestra en la cual ocurre"

Estos párrafos ilustran que un outlier es un concepto subjetivo trabajado primeramente como una inspección visual y desarrollado a través del tiempo teóricamente, por lo cual en la actualidad es ampliamente aceptado, entendido y utilizado por los estadísticos y personas que trabajan es estadística.

#### I.1.2 Efectos de los outliers

	<p align="center"><b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b></p>	<p>CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :7 DE 28 Fecha: 24-06-02</p>
---	--	---

Para un correcto y eficiente análisis de la información proveniente de comercio exterior es esencial determinar que registros son o no atípicos. Ya que su presencia distorsiona los estadísticos utilizados para el estudio del comercio exterior. De tal forma establecer outliers y con ellos los rangos es una labor básica y esencial para el mejoramiento de la calidad de información suministrada por el DANE.

### I.1.3 Influencia de los outliers en la estimación estadística

Conocida entonces la importancia de los outliers para el DANE, es primordial escoger entre los múltiples métodos para detectarlos "**el mejor**" para poder decir con seguridad antes de proceder con análisis estadísticos específicos cuales datos son o no outliers.

Establecer que datos son atípicos por lo tanto es una tarea delicada, ya que si no se detectan de forma correcta los rangos para las partidas arancelarias serán; o muy amplios permitiendo el paso de información discordante y contaminante a las bases de datos, o muy estrictos dejando un gran numero de información valiosa por fuera.

### I.1.4 Análisis de regresión

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables, para el desarrollo de este trabajo se considera solamente el modelo de regresión lineal simple. El modelo contiene un regresor  $X$  que tiene una relación lineal con la variable respuesta  $Y$ .

El modelo de regresión lineal simple es:

$$Y \equiv \beta_0 + \beta_1 X + \varepsilon$$

Donde el intercepto  $\beta_0$  y la pendiente  $\beta_1$  son constantes desconocidas y el error  $\varepsilon$  es un componente aleatorio. Los errores se asumen no correlacionados, con media cero y varianza  $\sigma^2$  los parámetros  $\beta_0$  y  $\beta_1$  son usualmente llamados **coeficientes de regresión**.

#### I.1.4.1 Estimación Mínimos Cuadrados de los Parámetros

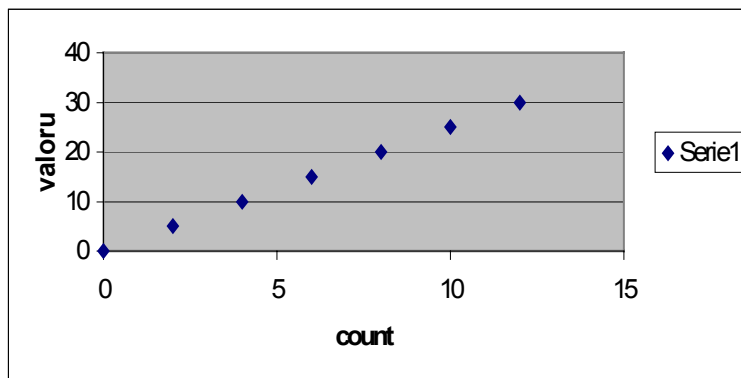
El método de mínimos cuadrados es usado para estimar los parámetros  $\beta_0$  y  $\beta_1$  y consiste fundamentalmente en la estimación de los parámetros de tal forma que la suma de cuadrados de las diferencias entre las observaciones  $Y_i$  y su estimación  $\hat{Y}_i$  sea mínima, es decir:



$$\sum \ell_i^2 \equiv \sum (y_i - \bar{y})^2 \equiv 0. \quad (1)$$

Por lo tanto la ecuación (1) va a permitir encontrar la línea de mejor ajuste para el modelo, es decir la línea que va a pasar lo más cerca posible del mayor número de puntos en un diagrama de dispersión.

### Ecuación de Regresión Vía mínimos Cuadrados



La estimación del coeficiente de regresión  $\beta_1$  es dada por:

$$\beta_1 \equiv \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

Y la estimación de  $\beta_0$  es:

$$\beta_0 \equiv \bar{y} - \beta_1 \bar{x}$$

El método de mínimos cuadrados proporciona los mejores estimadores lineales insesgados de  $\beta_0$  y  $\beta_1$ .

#### I.1.5 Análisis de residuales

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :9 DE 28 Fecha: 24-06-02
---	--	--

Una vez construido el modelo de regresión, se tiene que comprobar las hipótesis de linealidad, normalidad, homocedasticidad e independencia de los residuos. A continuación se presentara la información que proporcionan los residuos sobre estas hipótesis.

Los residuos minimocuadráticos vienen dados por:

$$\ell \equiv (I - V)Y \equiv (I - V)(X\beta + U) \equiv (I - V)U$$

Su esperanza es cero y su matriz de varianzas y covarianzas es:

$$\sum_e \equiv E[ee] \equiv \sigma^2(I - V)$$

es decir

$$\text{Var}(e_i) \equiv \sigma^2(1 - v_{i_i})$$

donde  $v_{i_i}$  es el término diagonal de la matriz  $v$  que mide la distancia entre el punto  $x_i$  y la media  $\bar{X}$ .

Para comparar los residuos entre sí suele ser más ilustrativo estandarizarlos. definiremos los residuos estandarizados  $r_i$  por:

$$r_i \equiv \frac{e_i}{\hat{s}_R \sqrt{1 - v_{i_i}}}$$

donde cada residuo se divide por una estimación de su desviación típica. Esta expresión numerador y denominador son dependientes, ya que  $e_i$  se utiliza en el cálculo de  $\hat{s}_R$ . Podemos solucionar este problema eliminando el punto  $(y_i, x_i)$  de los cálculos y estimando una regresión de los restantes  $n - 1$  datos. Sean  $\hat{\beta}_{(i)}$  y  $\hat{s}_{R(i)}^2$  los estimadores así obtenidos.

$$\hat{s}_{R(i)}^2 \equiv \frac{(n - k - 1)\hat{s}_R^2 - e_i^2 / (1 - v_{i_i})}{n - k - 2}$$

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :10 DE 28 Fecha: 24-06-02
---	--	---

donde  $e_i$  y  $v_i$  tienen la interpretación habitual. Observemos que, como el punto  $(Y_i, X_i)$  no interviene en el calculo de  $\hat{s}_{R(i)}^2$ , el residuo  $e_i$  es independiente de  $\hat{s}_{R(i)}^2$ . Se define el residuo estudentizado,  $\hat{t}_i$ , por:

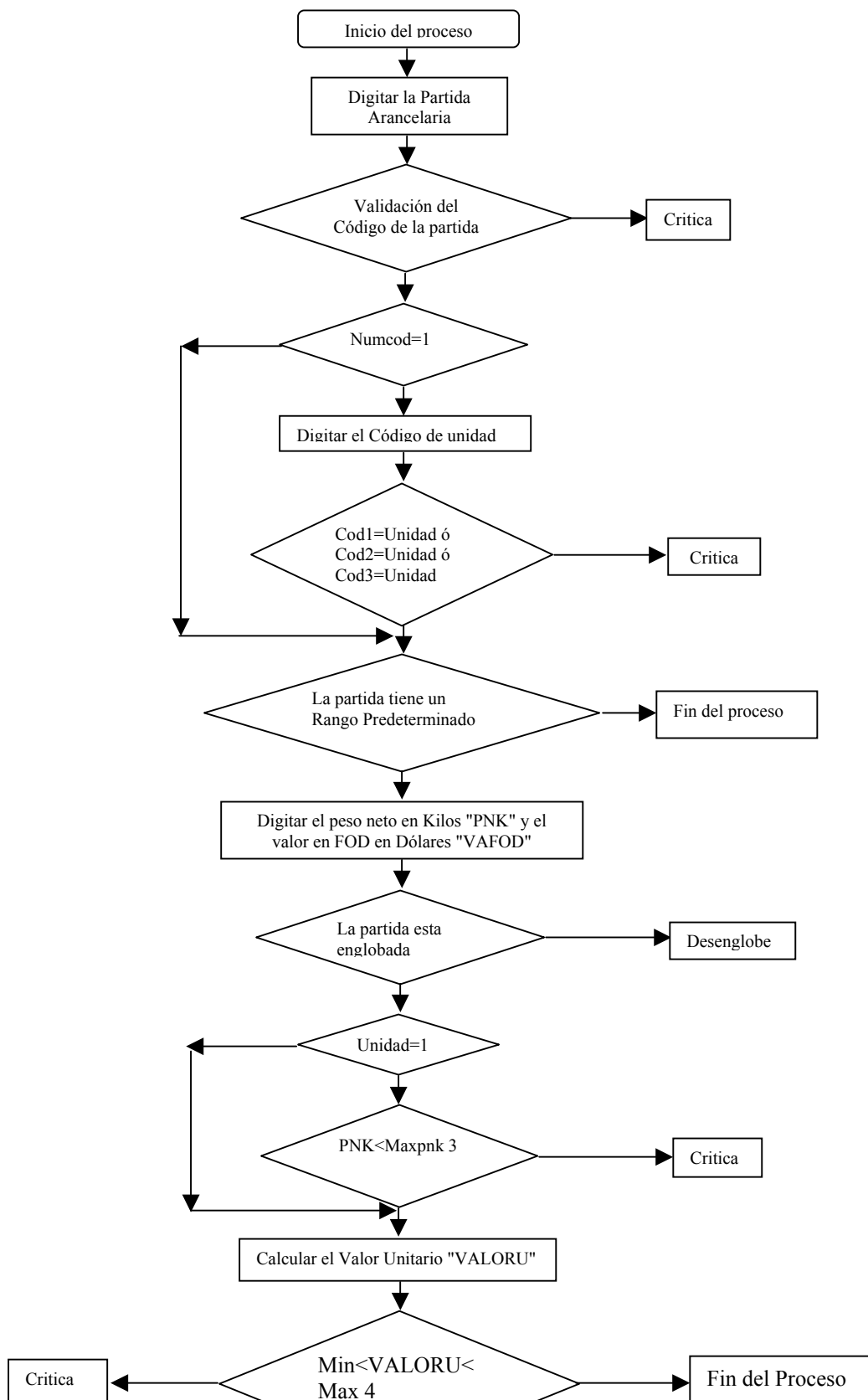
$$\hat{t}_i \equiv \frac{e_i}{\hat{s}_{R(i)}^2 \sqrt{1 - v_i}}$$

que tendrá una distribución de student con n-k-2 grados de libertad.

Estos tres residuos,  $e_i, r_i, \hat{t}_i$ , aportan información valiosa sobre los datos. Si n es grande, y si los datos no contienen valores extremos, el comportamiento de tres tipos de residuos es análogo. En otros casos, los residuos  $r_i$  y  $\hat{t}_i$  suelen ser más informativos para detectar deficiencias en el modelo.

## II. METODOLOGIA

En el siguiente esquema se presenta el diagrama de flujo de la metodología.



	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :12 DE 28 Fecha: 24-06-02
---	--	---

3 Maxpnk; es el número máximo de kilos admisibles.

4 Min y Max; son los valores mínimo y máximo del valor unitario

## II.1 Validación de la codificación de las posiciones arancelarias

La primera validación que se hace es la del código de la posición arancelaria. Para ello se tomó el archivo maestro de posiciones arancelarias que tiene División de Producción Estadística de la Dirección de Estadísticas Básicas del DANE. La aplicación de esta validación, consiste en determinar si el código de la posición arancelaria que está ingresando en el sistema de captura de la información de exportaciones, se encuentra entre los códigos del archivo maestro. En el caso en que la posición no sea válida, la grabación del registro debe terminar, y el documento debe ser remitido al crítico para que este determine el código correspondiente y válido.

## II.2 Validación de códigos de unidad

Siendo el interés principal de este estudio, el mejoramiento de la calidad de la información que ingresa periódicamente al DANE, desde sus primeras etapas hasta su análisis final se debe establecer de forma concreta los códigos de unidades que deben ser digitadas en el momento de capturar la información. Antes de proceder a estudiar el comportamiento de las variables peso neto en kilos y valor fod en dólares, se hace necesario determinar cuales son los códigos de unidad admisibles para cada una de las partidas arancelarias. Para este fin se conformo un archivo maestro el cual contiene para cada posición arancelaria hasta tres códigos posibles de unidad.

Cuadro 1. Códigos de Unidad Comercial de Medida, según resolución de la DIAN, utilizados a partir de enero de 1995.

Según DANE	Según resolución	Según DIAN	Descripción
66	66	BLL	Barril (Petroleo)
12	87	CEN	Centenar
15	15	DZN	Docena
21	21	DPR	Docena de pares
48	18	GRM	Gramo
23	16	GRO	Gruesa
20	20	SET	Juego
33	33	KGM	Kilogramos
	74	MWH	Megavatio hora
91	91	MTR	Metro
82	82	MTK	Metro cuadrado
67	67	MTQ	Metro cubico
45	64	MGM	Miligramo
13	45	MIL	Millar
11	11	NAR	Numero de artículos
19	19	NPR	Número de pares

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :13 DE 28 Fecha: 24-06-02
---	--	---

Según DANE	Según resolución	Según DIAN	Descripción
11	89	PCE	Pieza*
53	97	QTE	Quilate
42	42	TNE	Tonelada métrica
61	61	LTR	Litro
15	79	DPC	Docena de piezas

Fuente: Resolución DIAN y cartilla de códigos del DANE

Resolución: 0115 del 15 de enero de 1996 de la DIAN

Resolución: 5494 del 11 de septiembre de 1996 DIAN

\*En la cartilla del DANE no existe código para esta unidad, actualmente se esta capturando como 11, correspondiente a código de unidad.

La aplicación de la regla de validación se realiza de la siguiente manera: para las partidas arancelarias que admiten tan solo un código de unidad, el sistema de captura les asigna automáticamente la unidad. Para las partidas que admiten una de dos o tres posibles unidades, el sistema pide al digitador que grabe la unidad de medida que viene en el documento de exportación. Si la unidad de medida digitada no concuerda con una de las dos o tres posibles unidades del archivo maestro, la digitación del documento termina, siendo este remitido al crítico para que este lo revise.

Las siguientes dos posiciones arancelarias no tienen códigos: XXXXX

Las implicaciones que traería un error de codificación es una errada interpretación en las unidades, por ejemplo no es lo mismo 100 galones 100filos, por lo tanto digitar o codificar mal el código de unidad implica un degeneramiento de la información.

### II.3 Desenglobe del peso en kilos brutos o del valor en fob dólares

En el movimiento de comercio exterior, tanto de importaciones como de exportaciones, se presenta el caso de un documento arancelario que contiene más de una posición, el número de kilos de cada una de ellas no venga especificado, trayendo el documento tan sólo el número total de kilos.

Este mismo caso se presenta para valor de la mercancía, la cual viene englobada en un monto total, sin especificar el valor de cada posición arancelaria. En estos casos, o bien se conoce el valor de la mercancía de cada posición arancelaria, desconociéndose el número de kilos, o bien se conoce el número de kilos y se desconoce el valor de la mercancía.

La metodología aquí pretende resolver técnicamente el anterior problema, buscando estimar los valores desconocidos, ya sea de kilos o del monto de la mercancía, utilizando la información histórica del movimiento del comercio exterior. Aquí haremos referencia al caso de las exportaciones, teniendo en cuenta que esta metodología será implementada en el nuevo sistema de captura de la información de exportaciones.

### II.3.1 Metodología para estimación de kilos

Se tiene el siguiente cuadro de información:

Posición Arancelaria	Kilos netos	Valor Fob Dólares
1	?	$FOD_1$
2	?	$FOD_2$
...	...	...
N	?	$FOD_n$
Total	$K = \sum k_i$	$FOD = \sum FOD_i$

Se tiene que el valor total de la mercancía de una posición arancelaria resulta de multiplicar su valor unitario por el número total de kilos:

$$FOD_i \equiv VU_i * k_i$$


Donde  $FOD_i$  es el valor total en dólares de la mercancía;  $VU_i$  es el valor por kilo de mercancía y  $k_i$  es el número de kilos de posición arancelaria.

Teniendo en cuenta que tan sólo contamos con el valor  $FOD_i$  y desconocemos el número de kilos  $K_i$ , podemos utilizar el valor unitario  $VU_i^H$  obtenido del mes inmediatamente anterior, o de trimestre anterior o de otro período anterior, pero de la misma posición arancelaria, para estimar el número de kilos de la posición.

$$K_i^e \equiv FOD_i / VU_i^H$$

Donde  $K_i^e$  es el número de kilos estimado;  $FOD_i$  es el valor de la mercancía en dólares y  $VU_i^H$  es el valor unitario de la mercancía obtenido del archivo histórico.

Teniendo en cuenta que el  $VU_i^H$  es un valor histórico al utilizarlo directamente la suma de los kilos no coincidiría con el número total de kilos que reporta el documento, es necesario utilizar un factor de relajamiento que permita que estos coincidan. El factor de relajamiento lo denominaremos  $\alpha$ , el cual al multiplicarlo por el coeficiente de variación  $CV_i^H$  obtenido del mes inmediatamente anterior o de trimestre anterior o

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :15 DE 28 Fecha: 24-06-02
---	--	---

de otro período anterior, pero de la misma posición arancelaria, permitirá que la suma de los kilos estimados para cada posición, sea igual al total de kilos del documento.

$$FOD_i \equiv (VU_i^H * \alpha * CV_i^H) * K_i^e$$

$$K_i^e \equiv \frac{FOD_i}{(VU_i^H * \alpha * CV_i^H)}$$

Ahora bien, si aplicamos sumatoria a juntos lados obtenemos:

$$K \equiv \sum K_i^e \equiv \sum \frac{FOD_i}{(VU_i^H * \alpha * CV_i^H)}$$

Desarrollando la sumatoria se tiene:

$$K \equiv \frac{1}{\alpha} \left( \frac{FOD_1}{VU_1^H * CV_1^H} + \frac{FOD_2}{VU_2^H * CV_2^H} + \dots + \frac{FOD_n}{VU_n^H * CV_n^H} \right)$$

Despejando  $\alpha$  se tiene:

$$\alpha \equiv \frac{1}{K} \left( \frac{FOD_1}{VU_1^H * CV_1^H} + \frac{FOD_2}{VU_2^H * CV_2^H} + \dots + \frac{FOD_n}{VU_n^H * CV_n^H} \right)$$

$$\alpha^e \equiv \frac{1}{K} \sum \frac{FOD_i}{(VU_i^H * CV_i^H)}$$

Finalmente el número de kilos estimado para cada posición arancelaria resulta así:

$$K_i^e \equiv \frac{FOD_i}{VU_i^H * \alpha^e * CV_i^H}$$

### II.3.2 Metodología para estimación de Fob Dólares



	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :16 DE 28 Fecha: 24-06-02
---	--	---

Se tiene el siguiente cuadro de información:

Posición arancelaria	Kilos netos	Valor
1	$K_1$	
2	$K_2$	
.	.	
N	$K_n$	
Total	$K \equiv \sum K_i$	$FOD \equiv \sum FOD_i$

Se tiene que el valor total de la mercancía de una posición arancelaria resulta de multiplicar su valor unitario por el número total de kilos:

$$FOD_i \equiv V_{ui} * K_i$$

Donde  $FOD_i$  es el valor total en dólares de la mercancía;  $V_{ui}$  es el valor por kilo de mercancía y  $K_i$  es el número de kilos de la posición arancelaria.


Teniendo en cuenta que tan sólo contamos con el número de kilos por posición arancelaria  $K_i$  y desconocemos el valor  $FOD_i$ , podemos utilizar un valor unitario  $VU_i^H$  obteniendo el mes inmediatamente anterior, o de trimestre o de otro período anterior, pero de la misma posición arancelaria, para estimar el valor de la mercancía.

$$FOD_i^e \equiv \frac{K_i}{VU_i^H}$$

Donde  $FOD_i^e$  es el valor de dólares estimado para la posición  $i$ ;  $k_i$  es el número de kilos y  $VU_i^H$  es el valor unitario de la mercancía obtenido del archivo histórico.

Teniendo en cuenta que el  $VU_i^H$  es un valor histórico, al utilizarlo directamente la suma de los dólares de cada posición arancelaria no coincidiría con el valor total en dólares que reporta el documento, es necesario utilizar un factor de relajamiento que permita que ambas sumas coincidan. El factor de relajamiento lo denominaremos  $\beta$ , el cual multiplicado por el coeficiente de variación  $CV_i^H$  obtenido del mes inmediatamente anterior, o de trimestre anterior o de otro período anterior, pero de la misma posición arancelaria, permitirá que la suma de los dólares estimados para cada posición, sea igual al total de dólares del documento.

$$FOD_i^e \equiv (VU_i^H * \beta * CV_i^H) * K_i$$

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :17 DE 28 Fecha: 24-06-02
---	--	---

Ahora bien si aplicamos sumatoria a juntos lados obtenemos:

$$FOD \equiv \sum FOD_i^e \equiv (VU_i^H * \beta * CV_i^H) * Ki$$

Desarrollando la sumatoria se tiene:

$$FOD \equiv \beta * (VU_1^H * CV_1^H + VU_2^H * CV_2^H + \dots + VU_n^H * CV_n^H)$$

Despejando  $\beta$  se tiene:

$$\beta^e \equiv \frac{FOD}{(VU_1^H * CV_1^H + VU_2^H * CV_2^H + \dots + VU_n^H * CV_n^H)}$$

Finalmente el número de kilos estimado para cada posición arancelaria resulta así:

$$FOD_i^e \equiv (VU_i^H * \beta^e * CV_i^H) * Ki$$

Los parámetros a utilizar  $(VU^H, CV^H)$  están contenidos en el archivo maestro que se entrega con la metodología. Sin embargo, ésta presenta una limitación. Para las posiciones arancelarias que no tienen movimiento en el año 1996 no se puede calcular los parámetros, no siendo posible entonces la aplicación de la metodología.

**II.4 Valor de número máximo de kilos para posiciones arancelarias con unidad de medida número de unidades.**

Esta validación es solamente utilizada para las posiciones arancelarias cuya unidad de medida viene dada en número de unidades. El objetivo perseguido es el controlar que el número de kilos por unidad exportada sea coherente con el tipo de mercancía. Por ejemplo, si se están exportando asnos, es muy poco probable que un asno pese 20.000 kilos.

Si el número de kilos por unidad excede el valor contenido en la tabla maestra, la grabación del registro se debe terminar, pasando el documento al crítico. Si el crítico verifica que el dato es correcto, se sigue con la grabación, quedando marcado el registro como con un indicador de registro atípico en el número de kilos por unidad.

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :18 DE 28 Fecha: 24-06-02
---	--	---

La metodología estadística utilizada para el establecimiento del número máximo de kilos admisibles es similar a la utilizada en la sección II.5.2.

## II.5 Rango del valor unitario en dólares

Como control final tanto del valor de la mercancía, como del número de kilos netos, se elaboró un rango admisible para el valor unitario. Recordemos que el valor unitario es el resultado de dividir el valor total en dólares de la mercancía por el número total de kilos netos. Para la determinación de los rangos admisibles se elaboraron dos metodologías, cuyas descripciones se presentan a continuación.

Uno de los mayores problemas que se enfrenta en la elaboración de las índices de precios, es la purificación de los valores unitarios existentes en las bases de datos. Lo que se busca básicamente es obtener valores unitarios que reflejen el nivel de los precios y que al comparar la evolución de los mismos, no se estén involucrando factores exógenos a la variación misma de precios. Dentro de los factores que pueden afectar los valores unitarios, dada una calidad y confiabilidad en la información de los mismos, están los cambios de estructuras al interior de las partidas arancelarias, debido a cambios en las especificaciones técnicas de los artículos que la componen. Otros factores, son la calidad de los artículos y el país de procedencia para el caso de las importaciones.

En esta etapa de la metodología se construyen criterios para la limpieza de las partidas arancelarias, buscando obtener precios unitarios que estén en lo posible libres de las contaminaciones antes mencionadas.

Como el valor resulta de la división entre el valor total de la mercancía, expresado en cifras FOB en dólares, y el número total de kilos netos que pesa la mercancía, la calidad del dato del valor unitario depende de la calidad de los datos tanto del valor de la mercancía, como del número de kilos.

En este punto entran en juego muchos factores, que van desde la misma calidad de los datos, pasando por la especificaciones técnicas de la mercancía y el país de procedencia o destino, hasta la sub o sobrefacturación en que pueden estar incurriendo los exportadores, según sea el caso. Teniendo en cuenta que cada posición arancelaria se puede homologar muestralmente a un estrato de la población, lo que se busca es que el valor medio de los registros dentro de la posición se comporten lo más homogéneamente posible. Para que este objetivo se logre se deben descartar para el cálculo de los valores unitarios los registros que tengan este valor muy por debajo, o muy encima del valor medio de los valores unitarios de los registros dentro de la posición.

Al examinar el comportamiento se los VU se observa la existencia de valores extremadamente altos o bajos respecto a la mediana<sup>3</sup> de dicho valor. Muchos de estos valores pueden ser reales pero atípicos, u otros tal vez respondan a las

<sup>3</sup> Mediana: es el valor que es superado por el 50% de las observaciones y supera el restante 50%.

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :19 DE 28 Fecha: 24-06-02
---	--	---

sobrefacturaciones en el caso del "lavado" de dólares o a subfacturaciones en el caso de la evasión de impuestos. Teniendo en cuenta que tanto los uno como los otros se consideran atípicos, la metodología aplicada en este paso permite identificar los valores raros u outliers que sospechosamente parecen ser errados.

Es importante aclarar que esto no busca garantizar la depuración total al problema de sub o sobrefacturación que se presenta en los datos de comercio exterior, pero si aporta elementos validos sobre la depuración de la información contaminada por este tipo de problemas en el momento mismo de la grabación.

La metodología diseñada para este fin trata básicamente de conformar un rango de valores unitarios aceptables. Todos aquellos registros con valores por fuera de los límites de dicho rango se consideran atípicos, siendo revisados por el crítico.

Para la determinación de límites se puede hacer uso de diferentes harramientas estadísticas tales como los gráficos esquemáticos (BoxPlot) pertenecientes al análisis exploratorio de datos. Una segunda forma de determinar los limites consiste en estimar un número de desviaciones a partir de una media de posicionamiento tal como la mediana, formándose con éste un rango de valores aceptables. La tercera metodología considerada, consiste en realizar un análisis de residuos estudentizados. A continuación se presenta de las dos últimas metodologías.

#### II.5.1 Metodología de la regresión

La idea básicamente consiste en sumarle, para hallar la cota superior, o restarle para hallar la inferior, a una medida de posicionamiento central como la media o la mediana, un número determinado de medidas de dispersión tal como la desviación estándar<sup>4</sup> o una proporción del rango intercuartílico.

Para llevar a cabo esta tarea, en primera instancia se debe determinar cuál es la medida de posicionamiento central a utilizar. Teniendo en cuenta que los valores unitarios tienen en algunos casos mucha dispersión resultante especialmente de la presencia de valores extremos a la derecha, el valor de la media en estos casos se ve por la presencia de tales valores. Debido a lo anterior, no seria recomendable utilizar la media para los fines perseguidos, siendo la mediana una medida más robusta. Debido a que ésta, no se ve afectada por la presencia de valores extremos.

Seleccionada la mediana como medida de posicionamiento central, se pasa a la elaboración del modelo que permita la estimación de los límites del rango. Para ello se tratan las posiciones en tres grupos. El primero conformado por las posiciones que tienen un coeficiente de variación menor o igual a 200; el segundo por las posiciones que tienen coeficiente de variación mayor que 200 pero menor o igual a 300; y el tercero conformado por las posiciones con coeficiente de variación mayor a 300.

<sup>4</sup> Desviación estándar es la raíz cuadrada de la varianza.

La idea es determinar en cada caso el número de desviaciones estándar existentes entre la mediana y el valor de un percentil determinado. Como a medida que aumenta el coeficiente de variación, se tiene que algunos de los valores unitarios también tiende a estar mucho más alejado de la mediana, entonces se utilizan tres criterios diferentes para determinar los percentiles.

Cuadro No. 2 Criterios utilizados en el calculo del número de desviaciones estándar.

Grupo	Coeficiente de variación	Percentil limite inferior	Percentil límite superior
1	Hasta 200	2	98
2	De 200 a 300	4	96
3	300 u más	6	94

Con base en los anteriores criterios se calcula el número de las desviaciones estándar tanto para el límite superior como para el inferior de cada grupo conformado por posiciones arancelarias, utilizando las siguientes formulas:

$$STDPIN = (MEDIANA - PLI) / STD$$

$$STDPSU = (MEDIANA + PLS) / STD$$

Donde STDPIN es el número de desviaciones estándar entre la mediana y el percentil inferior, mediana es el valor de la mediana del valor unitario de cada posición, PLI es el valor del percentil del limite inferior de la posición (calculada a partir de los registros) y STD es la desviación estándar de los registros dentro de la posición. STDPSU es el número de desviaciones estándar entre la mediana y el percentil superior, PLS es el valor del percentil del limite superior de la posición (calculada a partir de los registros).\*

Con lo anterior tenemos los valores observados del número de desviaciones estándar según la regla establecida a priori. El paso a seguir es establecer cual es el número de desviaciones estándar a aplicar a cada posición. para ello es necesario determinar algunas características de la distribución de los

Valores unitarios que sirvan de referente para tal decisión. Teniendo en cuenta que estamos haciendo uso de dos medidas: de posicionamiento y de dispersión, utilizaremos un modelo de regresión para estimar el número de desviaciones que se

$$* S \equiv \sqrt{\frac{\sum W_i (X_i - \bar{X})^2}{\sum W_i}}$$

Donde S es la desviación estándar,  $\bar{X}$  es la media aritmética,  $X_i$  es cada valor de la variable X y  $W_i$  es el peso ponderado de cada valor de la variable X.

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :21 DE 28 Fecha: 24-06-02
---	--	---

aplicarán a cada posición, tomando como predictores el valor de la estadística apuntamiento<sup>5</sup> y Kurtosis<sup>6</sup> Los modelos de regresión a utilizar son los siguientes:

$$STDPIN_1 \equiv \beta_0 + \beta_1 * KURTOSIS_1 + \beta_2 * APUNTAMIENTO_1 + E_1$$

$$STDPSU_1 \equiv \beta_0 + \beta_1 * KURTOSIS_1 + \beta_2 * APUNTAMIENTO_1 + E_1$$

donde E es el error,  $\beta_0, \beta_1, \beta_2$  son los parámetros a estimar.

Para determinar los límites finales a utilizar se emplean las siguientes fórmulas:

$$MINI \equiv MEDIANA - (STDPINE) * STD$$

$$MAXI \equiv MEDIANA + (STDPSUE) * STD$$

Donde MINI es el rango inferior y MAXI es el rango superior, STDPINE es el número de desviaciones estándar estimado para el límite inferior y STDPSUE es el número de desviaciones estándar estimado para el límite superior.

Para todos los registros, cuyo valor unitario esté por fuera del anterior rango, se debe terminar la grabación y mandar los documentos al crítico, para que este determine si es posible que existan errores que puedan ser solucionados por el.

## II.5.2 Metodología de los Residuales Estudentizados

Otra forma de establecer los rangos que filtran la información proveniente de comercio exterior se realiza a partir del concepto de outliers. Ya que si determinamos para los registros de una partida arancelaria desde qué punto un valor comienza a ser outliers también determinamos hasta qué valor no lo es, esto hace posible establecer el valor máximo y mínimo admisible.

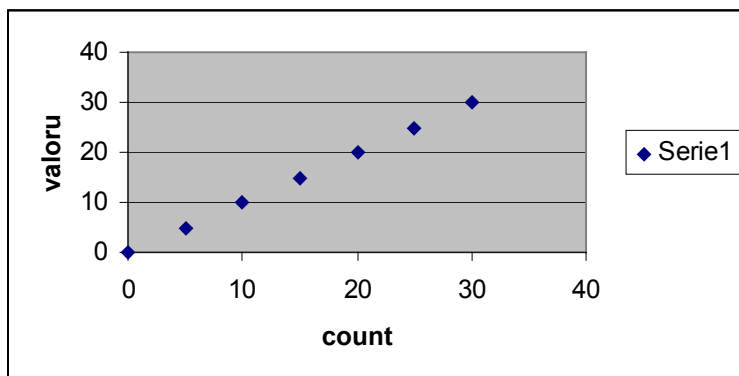
### Graficos de Dispersión

<sup>5</sup> Apuntamiento: medida estadística que se refiere a la altura que tiene una distribución probabilística.

Apuntamiento  $\equiv \sum Z_i^3 n / (n-1)(n-2)$ , donde n es el número de observaciones y z es la variable estandarizada.  $Z_i \equiv (X_i - \bar{X}) / S$ , con  $\bar{X}$  la media aritmética,  $X_i$  cada valor de la variable X y S la desviación estándar.

<sup>6</sup> Kurtosis: medida estadística que mide la forma aplanada que tiene una distribución probabilística.

Kurtosis  $\equiv \sum_i^4 n(n-1)/(n-1)(n-2)(n-3) - 3(n-1)^2 / (n-2)(n-3)$ , donde n es el número de observaciones y z es la variable x estandarizada.  $Z_i \equiv (x_i - \bar{x}) / s$ , con  $\bar{x}$  la media aritmética,  $x_i$  cada valor de la variable X y S la desviación estándar.



Un outlier es un concepto subjetivo postdata

Históricamente los métodos objetivos para trabajar con outliers fueron empleados únicamente después de una inspección visual de los datos. Para fijar ideas, supongamos que las siguientes observaciones son provenientes de realizaciones independientes de una población normal común.

-1.64	-133	-1.10	-0.57	-0.27
1.04	1.56	1.84	2.04	4.99

Cual de estas observaciones se puede llamar Outlier? La respuesta es que podemos reservarla al juzgamiento bajo una prueba o alguna medida objetiva que haya sido construida, pero la decisión para emplear un criterio objetivo esta frecuentemente basada en inspecciones iniciales de los datos. En este caso, la observación 4.99 se ve más grande y se puede preguntar acerca de la validez de la segunda observación más grande 2.04.

Una forma de percibir un outlier depende del método de presentación de los datos (aleatoria, ordenadamente o gráficamente), en la experiencia, y en la escala de los datos, observaciones extremas tienden a aparecer más discretamente en tanto la escala se incrementa.

En conjuntos grandes de datos o análisis más complicados una inspección visual de los datos puede ser imposible, de esta forma es necesario aplicar algún tipo de criterio objetivo, más allá de una inspección visual de los datos, muchos de estos criterios son basados en modelos y de esta forma algunas decisiones acerca de la naturaleza y frecuencia de los outliers es necesaria antes de tomar alguna decisión. la subjetividad es introducida en el proceso de modelación y lo que se desea es que el criterio resultante sea aplicado necesariamente en una rutina y de manera objetiva.

Como nuestro objetivo es filtrar la información discordante de exportaciones, es decir determinar los valores contaminantes que pueden tomar la variable peso neto en kilos

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :23 DE 28 Fecha: 24-06-02
---	--	---

(PNK) y el valor FOD en dólares (VAFOD) debidos a errores de registro o digitación de la información y que alteran todo posible análisis de datos de las partidas arancelarias, se deben establecer unos valores máximos y mínimos para cada una de las variables.

Ahora bien, determinar cuales valores son realmente atípicos o no en la variable PNK o VAFOD es una labor compleja debido a que se cuenta con información de grandes y pequeños exportadores, por lo cual es muy probable determinar como outlier los valores de las variables PNK o VAFOD correspondientes a una empresa grande, cuando en realidad no lo son. Si además de esto añadimos el hecho de que una empresa pueda realizar en el año solo una exportación de gran volumen, de fijo detectaremos un outlier en los kilos netos (PNK) o en el valor FOD (VAFOD) cuando en realidad es simplemente el reflejo de una observación poco frecuente pero admisible. Por este motivo es mejor establecer rangos para la variable valor unitario ya que esta variable no se encuentra afectada por el volumen de exportación.

#### II.5.2.1 Algoritmo

El proceso para determinar el rango de la variable valor unitario (VALORU) correspondiente a una partida arancelaria consiste como se menciono anteriormente en la detección de los outliers para que de forma iterativa se pueda establecer el valor máximo admisible.

El valor unitario se define como la razón entre el valor FOD y el número de artículos exportados y representa el costo que tiene en dólares de exportar un artículo.

$$VALORU \equiv Vafod / Pnk$$

Luego de creada la variable que contiene los valores unitarios se ordena de forma ascendente, posteriormente se crea una variable índice que indique la posición de los registros en el conjunto de datos. Se realiza una Regresión entre la variable índice y valoru.

$$valu \equiv \beta_0 + \beta_1 indice$$

En base a los resultados obtenidos en el análisis de regresión se determinan los residuos estudentizados, se establece el valor del estadístico T para el conjunto de datos (con un nivel de significancia del 92%), que como se menciono anteriormente nos determina el valor máximo que deben tomar los residuos estudentizados. Por tal motivo cualquier valor de los residuos estudentizados calculados en la regresión que sobrepase el estadístico T se puede considerar discordante, lo que nos indica que el valor asociado a este residuo estudentizado es un outlier. En este punto puede suceder que se encuentren o no outliers.



	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :24 DE 28 Fecha: 24-06-02
---	--	---

Si no se encuentran outlier se le adiciona al valor máximo de la variable VALORU un porcentaje de la desviación estándar de la variable en estudio, este porcentaje se determina según el coeficiente de variación de la variable, bajo el siguiente criterio; Si el coeficiente de variación es menor que 200\*, se le adiciona un décima\* parte de la desviación estándar de la variable. Si el coeficiente se encuentra entre 200\* y 300\* se le adiciona una octava\* parte de la desviación y finalmente si el coeficiente es mayor que 300 se le adiciona un sexta\* parte de la desviación.

$$\hat{X}_n \equiv X_n + \frac{(std)*10}{100} \text{ si el } cv < 200.$$

$$\hat{X}_n \equiv X_n + \frac{(std)*8}{100} \text{ si el } 200 < cv < 300.$$

$$\hat{X}_n \equiv X_n + \frac{(std)*6}{100} \text{ si el } cv > 300.$$

Los valores utilizados como criterios (\*) fueron establecidos en base a un estudio preliminar de la variable VALORU para cada una de las partidas existentes en el estudio.

Si por el contrario se encuentran outliers al comparar los residuos estudentizados con el estadístico T, se procede establecer el último valor de la variable ( VALORU) que no es outlier, supongamos que su posición dentro del conjunto de datos sea “n” y el primer outlier encontrado, es decir “n+1”.

Se incrementa el último valor en un determinado porcentaje d la diferencia entre n y n+1, el porcentaje se establece de acuerdo al valor del coeficiente de variación así:

Si el valor es menor que 200 se incrementa el último valor en un 25% de la diferencia, si se encuentra entre 200 y 300 se incrementa en un 50% y finalmente si es mayor que 300 el incremento es del 75% de la diferencia.

$$\hat{X}_n \equiv X_n + \frac{(\Delta_{n,n+1})*25}{100} \text{ si el } cv < 200.$$

$$\hat{X}_n \equiv X_n + \frac{(\Delta_{n,n+1})*50}{100} \text{ si el } 200 < cv < 300$$

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :25 DE 28 Fecha: 24-06-02
---	--	---

$$\hat{X}_n \equiv X_n + \frac{(\Delta_{n,n+1}) * 75}{100} \text{ si el cv} > 300$$

donde  $\Delta_{n,n+1}$  es la diferencia entre el valor unitario de la posición n y n+1.

El proceso se repite de forma iterativa, de tal manera que al cabo de algunas iteraciones se establezca un valor para la variable VALORU para el cual su residuo estudentizado luego de realizar el análisis de regresión sea aproximadamente igual al estadístico T.

$$\hat{T}_n \cong T_{n-k-2}^{0.92}$$

donde  $\hat{T}_n$  es el residuo estudentizado asociado al valor unitario  $\hat{X}_n$ .

En conclusión este procedimiento lleva a la obtención de un valor de la variable valoru que sea el máximo valor admisible o permisible dentro del contexto de los datos de la variable.

Esto quiere decir que cualquier valor que sea inferior a  $\hat{X}_n$  no es un outlier y viceversa.

$$\text{Si } X_j > \hat{X}_n \Rightarrow X_j \text{ es un outlier } \forall J$$

Por lo tanto el intervalo queda determinado por:


$$(0, \hat{X}_n)$$

La convergencia de la estimación del valor máximo es relativamente rápida, según simulaciones hechas al programa, por lo tanto el número de iteraciones para efectos de tiempo y programación fue determinado como 5.

### II.5.3 Comparación de las dos metodologías

Las dos metodologías expuestas presenta similares resultados en los rangos estimados para los valores unitarios por lo que escoger una u otra no es cosa fácil. A continuación se expondrá las ventajas y desventajas de cada una de las metodologías.

**Cuadro 3. Distribución de las exportaciones en dólares según método de control de los valores unitarios.**

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :26 DE 28 Fecha: 24-06-02
---	--	---

Método de regresión												
Regresión	1	2	3	4	5	6	7	8	9	10	11	12total
Fuera de rango	5,42	5,27	6,02	7,28	7,13	2,95	4,97	4,68	7,72	10,7	6,34	10,55 6,71
Dentro de rango	94,58	94,7	93,98	92,72	92,87	97,05	95,03	95,32	92,28	89,3	93,66	89,45 93,29
Método de los residuos estudentizados												
Fuera de rango	5,52	5,1	5,12	3,37	5,92	4,9	3,85	3,08	1,98	7,59	2,03	9,63 4,86
Dentro de rango	94,48	94,9	94,88	96,63	94,08	95,1	96,15	96,92	98,02	92,41	97,97	90,37 95,14

#### II.5.3.1 Ventajas de la metodología de regresión

#### II.5.3.2 Ventajas de la metodología de los residuos estudentizados

Tal vez una de las principales ventajas de esta metodología es la capacidad de poder establecer el rango para una partida arancelaria utilizando solamente información de la respectiva partida, esto quiere decir que no se ve influencia de ninguna forma por los comportamientos de las demás partidas arancelarias.


Otra de las ventajas de la metodología es la eficiencia con la cual se estiman los valores máximos que puede tomar las partidas arancelarias, ya que los valores estimados concuerdan perfectamente con la impresión visual generada por la estructura de los datos, al realizar un análisis exploratorio (grafico) de los datos.

#### II.5.4 Periodo considerado en el estudio

Las bases de datos utilizadas para determinar los rangos son las referentes al año de 1996, ya que el diseño de la metodología esta proyectada para ser actualizables, esto quiere decir que se pueden seguir obteniendo rangos actualizados a partir de las bases de datos del año anterior. Se podría pensar en tomar las bases de todos los años anteriores a manera de base histórica, pero aunque esto fuera lo óptimo para muchos casos, para este no, ya que las exportaciones están afectadas por el tiempo y en particular los valores de exportación por la inflación. También se podría pensar en tomar sólo el último trimestre del año 96, pero la poca información de este período hace imposible aplicar las metodologías diseñadas. En conclusión tomar todo el período de exportaciones de 1996 es la mejor opción.

#### II.5.5 Información necesaria para la implementación de los metodologías

Para estimar los valores máximos y mínimos que deben tomar los valores unitarios se deben tener las variables peso neto en kilos y valor Fod en dólares para cada partida

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET- EXPO-01 VERSIÓN : 01 Página :27 DE 28 Fecha: 24-06-02
---	--	---

arancelaria, ya que como se mencionó anteriormente la variable valor unitario es la razón entre estas dos variables.

El problema de información surge realmente en el numero de registros de cada una de las partidas arancelarias, ya que se requiere un mínimo para poder obtener estadísticos válidos y sobre todos para poder realizar el análisis de regresión, elemento básico para la construcción de los rangos para los valores unitarios de comercio exterior, el número mínimo de registros se estableció en 7 ya que con este número se cubre un gran número de partidas arancelarias y la estimación de los estadísticos y los parámetros de regresión siguen siendo aceptables. Un requerimiento adicional para la metodología de regresión es tener un mínimo de cinco diferentes tipos de partidas por capítulo, los capítulos se determinan por los primeros números del registro Nandina, por lo tanto los diferentes tipos de partidas para un determinado capítulo se establecen por las diferencias en los últimos ocho números del registro Nandina.

#### II.5.6 Actualización de los archivos maestros

La actualización de los archivos maestros está referida a dos niveles. El primero corresponde a la actualización que se le debe dar a los códigos de las partidas arancelarias, lo mismo que a las unidades de medida. Esta actualización corresponde a apertura, cierre o creación de nuevas posiciones arancelarias. Debido a que la legislación de comercio exterior puede cambiar en cualquier momento del tiempo, este nivel de actualización debe realizarse en el momento en el cual surjan los cambios en la nomenclatura.

Buscando rapidez en la actualización de nomenclatura (Nandina y unidades de medida), sería recomendable que frente a esta tarea este el grupo de producción estadística de comercio exterior, el cual tendría la responsabilidad de estar pendiente de las sucesivas actualizaciones que sean necesarias, lo mismo que de entregar los archivos maestros con las actualizaciones, a las respectivas Direcciones Territoriales del DANE en las cuales se graba la información.

El segundo nivel de actualización, comprende el de los rangos de valores unitarios y número máximo de kilos. Esta actividad debe hacerse al inicio de cada año, una vez se cuente con el movimiento completo del año anterior. Para ello, se deben correr los programas en SAS que permiten dicha actualización. De nuevo, sería recomendable que esta llevad a cabo por el grupo de producción estadística de comercio exterior, el cual tendría la responsabilidad de producir los nuevos rangos, lo mismo que de entregar los archivos maestros con las actualización a las respectivas Direcciones Territoriales del DANE en las cuales se graba la información.

## IV.RESULTADOS

Las dos metodologías propuestas, en forma general, son igualmente eficientes en la detección de valores unitarios atípicos. Sin embrago se espera analizar en lo que resta del año, detalladamente las ventajas y desventajas de cada una de ellas. Para fines de

	<b>METODOLOGIA ESTADISTICA PARA EL CONTROL DE CALIDAD DE LA INFORMACION DE EXPORTACIONES</b>	CÓDIGO: LR-MET-EXPO-01 VERSIÓN : 01 Página :28 DE 28 Fecha: 24-06-02
---	--	---

prueba de la metodología, en el archivo que será entregado al grupo de sistemas de la Dirección de Estadísticas Básicas, se incluirán los rangos obtenidos con la metodología de los residuos estudentizados, no siendo por ello. La metodología de rangos definitiva para el sistema. Sin embargo este aspecto no afecta en nada el desarrollo del sistema de captura.

## **V.CONCLUSIONES Y RECOMENDACIONES**

Con el desarrollo de la metodología de control de calidad de los datos de exportaciones. Se está dando un paso gigantesco en lo referente a garantizar mayor confiabilidad de la información, lo mismo que en la automatización de los procesos de depuración de la misma. Sin embargo como toda metodología, la presente es susceptible de mejoramiento, evento que podrá madurarse con la utilización de la misma.

Un apoyo fundamental para el logro de calidad de los datos, es el montaje de un sistema de control de calidad de grabación. Para ello se recomienda anexarle al sistema de captura, un procedimiento que identifique una muestra de registros a los cuales se les va a realizar confrontación entre medio físico (formulario) y medio magnético (archivo del sistema). Para ello es necesario determinar un tamaño de muestra óptimo, que brinde por un lado alta confiabilidad en la inspección, pero que a su vez considere el esquema operativo con el cual se captura la información, es decir, se debe considerar que el tamaño de muestra no desborde la capacidad operativa disponible para el proceso de grabación. Otras determinaciones tienen que ver con el período de control, variables a controles, errores máximos admisibles, y en general, la forma de poner en marcha el control de calidad a la grabación.

## **BIBLIOGRAFÍA**

DIAN, 1995, “Estadísticas básicas de comercio exterior. Metodología y procedimientos”, Santafé de Bogotá, D.C. Colombia.

Moreno, Hernando (1997), “Metodología de los índices de Comercio Exterior para el PIB trimestral”, Mimeo de la Dirección de Estadísticas del DANE, Santafé de Bogotá.

SAS Institute, 1990, “SAS Language”, Care, NC, USA

SAS Institute, 1990, “SAS/STAT User’s Guide”, Volumen 1 y 2, Care, NC, USA.