SELECCIÓN DE LA METODOLOGIA PARA DETERMINAR ATIPICOS EN LAS BASES DE CALCULO DE UN INDICE DE COSTOS

Luz Adriana Hernández Vargas

FUNDACION UNIVERSITARIA LOS LIBERTADORES

ESPECIALIZACION EN ESTADISTICA APLICADA

BOGOTA D.C.

2015

SELECCIÓN DE LA METODOLOGIA PARA DETERMINAR ATIPICOS EN LAS BASES DE CALCULO DE UN INDICE DE COSTOS

Luz Adriana Hernández Vargas

Trabajo de Investigación Aplicada, para optar el título de
Especialista en Estadística Aplicada

FUNDACION UNIVERSITARIA LOS LIBERTADORES

ESPECIALIZACION EN ESTADISTICA APLICADA

BOGOTA D.C.

2015

Nota de aceptación

_____

_____

_____

_____

_____

_____

_____
Presidente del Jurado

_____
Jurado

_____
Jurado

Bogotá _____ de Octubre _____

AGRADECIMIENTOS

Expreso mi aprecio, agradecimiento y admiración por mis compañeros, sin cuyo esfuerzo, tiempo y ejemplo no hubiese sido posible culminar este trabajo. La paciencia de mi familia, así como, el conocimiento y orientación del cuerpo docente.

Gracias por cada bendición recibida

GLOSARIO

Atípico

Se refiere a un dato que es diferente a los demás registros de la muestra, requiriendo la revisión de su calidad con el fin de asegurar que las conclusiones obtenidas de dicha muestra, correspondan a la realidad estudiada.

Se utiliza también el nombre Outliner dentro de este trabajo para describir la presencia de un dato atípico.

Cuartil

Se refiere a los tres valores que dividen el conjunto de datos de una base (muestra), previo ordenamiento en cuatro partes porcentualmente iguales.

Diagrama de cajas y bigotes

Representación visual que permite identificar los cuartiles de un conjunto de datos, así como sus atípicos

Novedad Técnica

Codificación asignada a un registro dentro de la base de precios del índice, que permite la identificación de los bienes y servicios a los cuales se les realiza seguimiento, clasificándolos entre: variedades que por su calidad se pueden determinar como sustitutos perfectos, que no son comparables, y registros nuevos (registros que en general imposibilitan el cálculo de variación de precios porque no tienen precio anterior) (DANE, 2013)

Precios comparables

Dentro de la metodología de los índices de precios y costos, el cálculo de la variación de precios solo es posible cuando los bienes y servicios seleccionados dentro de la canasta de seguimiento, tienen cualidades o calidades comparables. En el caso en

que las calidades de los bienes o referencias no sean comparables, es imposible realizar la variación de precios y el registro sale de la base. (DANE, 2013)

Rango intercuartilico

Diferencia entre el tercer y primer cuartil de una distribución

Relativo de precios

Se refiere al relativo (relación) entre el precio actual y el precio anterior para un artículo, en una fuente específica.

RESUMEN

Este trabajo presenta los resultados de la aplicación de diferentes metodologías que permiten seleccionar los valores atípicos, en las bases de datos usadas para calcular los resultados en un índice de precios. Dentro de las metodologías seleccionadas se incluye una variante del algoritmo de Tukey, aquella que permite la construcción del grafico de cajas y bigotes, la definida como de “distancia media” y la prueba de Dixon.

Dado que fue posible el análisis de las metodologías propuestas para solo tres periodos diferentes, y subsecuentes a la implantación de una actualización en la canasta de recolección aplicada al indicador, se obtienen resultados no coincidentes, lo que determina la necesidad de continuar estudiando los resultados hallados para periodos posteriores en por lo menos un año más. Sin embargo, se denota claramente que el análisis de la totalidad de la base, a cargo del equipo temático, consume recursos innecesariamente ya que la magnitud de las modificaciones fruto de dicha verificación es muy baja (El máximo encontrado no supera el 2%)

De otro lado, la verificaron de las bases de datos sugirió la necesidad que los responsables del indicador analicen la periodicidad de toma de información para una parte importante de la canasta recolectada, dado que es muy posible que dicha periodicidad no se suscriba semestralmente, sino que sea solo necesaria cada año.

PALABRAS CLAVE

Algoritmo de Tukey, cuartil, valores atípicos, límite superior y límite inferior, mediana, media truncada, índice de precios, prueba de Dixon y diagrama de cajas y bigotes.

CONTENIDO

LISTADO DE GRAFICAS

LISTADO DE TABLAS

1. PLANTEAMIENTO DEL PROBLEMA

Dentro de las actividades previas en el proceso de cálculo de un índice de precios –y en general de cualquier proceso estadístico- se ubica la revisión de los datos a utilizar en la medición. La ejecución de dicha revisión debe atender varias consideraciones, entre las que se destacan la limitación de recursos, la oportunidad en la revisión, la necesidad de limitar el error humano dentro del proceso, así como su transparencia, de manera tal que la selección de los datos sometidos a la revisión, atienda criterios adecuadamente documentados, excluyendo en lo posible, consideraciones subjetivas de las personas involucradas en el proceso.

Actualmente, la producción de un índice de precios generado en una entidad encargada del tema, requiere la revisión de la información base que permite su cálculo; un proceso que hasta la fecha, implica la revisión de la totalidad de los registros colectados.

El panorama actual ha hecho necesaria la destinación de recurso humano para la verificación de aproximadamente 4000 registros en cada periodo operativo, que junto con el tiempo destinado al análisis de la información implican una carga operativa relevante al grupo encargado, así como el aumento de la probabilidad de error. Hasta el momento no se cuenta con una metodología que permita la detección de valores atípicos para su verificación prioritaria y el aseguramiento de la calidad de los datos sin exagerar el esfuerzo de revisión.

Dado lo anterior, se hace relevante la revisión de las metodologías para la detección de valores atípicos, especialmente referidas a los datos insumo en el cálculo de índices de precios y costos, de forma tal que la evaluación de las mismas pueda concluir en el mejor método aplicable en la detección de valores atípicos y la priorización de su revisión, mejorando los tiempos de entrega de la información para calculo, disminuyendo las cargas laborales al grupo encargado y documentando adecuadamente el proceso de manera tal, que sea fácilmente implementado por las personas involucradas y responsables.

Es posible partir de la experiencia internacional en el manejo de la problemática descrita: Los lineamientos que describen las mejores prácticas en el diseño de índices de precios (OIT; FMI, 2006), presentan una variante del algoritmo de Tukey, y la denominada metodología de “distancias medias”, pero también es posible acoger la conocida metodología de construcción de cajas y bigotes, especialmente aplicadas en índices de precios (Saìdi, 2005) y la prueba de Dixon

El presente trabajo describe los resultados obtenidos al aplicar la variante del logaritmo de Tukey, así como en el uso de la metodología de cajas y bigotes descrita típicamente en los libros de texto (Vena, 2014), la prueba de Dixon y finalmente, la metodología de determinación de atípicos calculando “distancias medias” (UNITED NATIONS, 2009). Los cuatro métodos son aplicados para las bases de datos disponibles: primer y segundo semestre de 2014, y primero de 2015; en cada escenario se confrontan los resultados de cada metodología de detección de atípicos, versus las detecciones

(modificaciones) realizadas al revisar cada uno de los registros contenidos en la base (método actual o grupo control). De esta manera es posible determinar la cantidad de registros que fueron detectados mediante uno u otro mecanismo, así como aquellos que requerían revisión (fueron detectados al revisar el 100% de la base), pero no fueron detectados por el método propuesto.

A partir de dichos resultados se concluye sobre el uso de una u otra metodología, calificándolas a partir de dos criterios jerarquizados: a) la más deseable será, aquella que permita detectar la mayor cantidad de registros que efectivamente requieren revisión, y b) que en segunda instancia, determine la menor cantidad de registros que finalmente no necesitan verificación[1] (No fue necesaria ningún modificación a la base de datos, después de ejecutado el análisis del 100% de la base).

De esta manera se propondría una metodología clara que permita definir los valores atípicos en las bases posteriores, y/o las recomendaciones para continuar analizando la información recolectada y la toma de decisión respecto de un método en un futuro cercano, todo lo anterior con el fin de disminuir los tiempos de entrega y cargas de trabajo del personal a cargo, manteniendo la calidad en los resultados.

[1] En este sentido la verificación hace referencia a la necesidad de modificar o excluir el registro que presenta el atípico.

2. MARCO DE REFERENCIA

El análisis de atípicos se sustenta en el análisis específico de los comportamientos de cada base revisada. Es posible utilizar el promedio y la mediana para determinar aquellos registros que se ubican en limites superiores e inferiores y que dado su comportamiento, ameritan su revisión por posibles errores de digitación, y de toma (recolección).

La selección de un método para determinar atípicos, depende de las características particulares de la muestra a analizar, y su aplicación permite disminuir el número de registros a revisar, mejorar los tiempos de entrega y documentar el método de forma tal que éste sea de fácil incorporación con funcionarios nuevos, sin depender de criterios subjetivos.

Un método continuamente referenciado para la ubicación de atípicos, dentro de la experiencia internacional de análisis de bases insumo en el cálculo de índices de precios y costos, implica el logaritmo de Tukey (OECD; FMI; OIT; NU; BM, 2006)

Esta metodología exige una revisión previa de la base de datos y en primera instancia excluye de la verificación, todos aquellos registros que describen una falta de variación. (Ya que en este caso se trata del análisis de información para el cálculo de variaciones de precios, la variable estudiada refiere al relativo de precios para los bienes y servicios seleccionados en una canasta representativa del índice trabajado, en donde el relativo es igual a la relación entre el precio actual y el precio anterior).

De otro lado, los lineamientos internacionales también presentan la metodología de “Distancias medias”, que permite el cálculo de los límites inferior y superior, para definir aquellos registros calificados como atípicos. Este método parte del análisis de cuartiles y el promedio de distancias presentes entre el tercer cuartil y mediana y entre la mediana y el primer cuartil.

Finalmente, se ubica el popular método de determinación de atípicos, definido en la gráfica de cajas y bigotes, así como la prueba de Dixon.

A continuación se describen los aspectos fundamentales de una y otra metodología, partiendo de un supuesto necesario para aplicarlas: las muestras observadas de variaciones de precios tienen una distribución normal.

Es importante considerar que la recolección de información a analizar, describe el comportamiento de una muestra, no del universo, por lo que se asume que la distribución de dichos registros en el tiempo, es normal.

Igualmente vale la pena destacar que el análisis realizado a los diferentes métodos hacen referencia a la verificación de atípicos para los relativos de precios (precio actual / precio anterior), no sobre el valor del precio como tal.

2.1. ALGORITMO DE TUKEY

El algoritmo de Tukey incorpora el uso de la media truncada y recorta la base inicial de trabajo en un 10%, predefiniendo como atípicos todos los registros allí contenidos.

El método de cálculo de los límites superior e inferior implica:

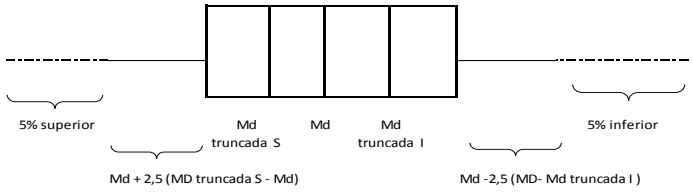
a. Ordenar los relativos de precios de mayor a menor, señalizando aquellos registros localizados en el 5% superior y 5% inferior (esta "base recortada", se cataloga a priori como casos atípicos).
b. La base de trabajo para la determinación del resto de atípicos se constituye por el restante 90% de registros, pero retirando además, todos los relativos de precio que señalan que no existió cambio en el mismo (relativo = 1)
c. Determinación de dos grupos de trabajo para la misma base: primera y segunda parte, definidas por medio de la media aritmética.
d. Calculo de la media aritmética para cada uno de los grupos: medias de las mitades: media superior y media inferior o medias truncadas.
e. Calculo de los limites superior e inferior a partir de:

   Límite superior: Media + 2,5 (media superior – media)
   Límite inferior: Media – 2,5 (media – media inferior)

Parte del éxito del uso de la media, consiste en retirar previamente los relativos de precio que no registran movimiento (relativo =1) (OECD; FMI; OIT; NU; BM, 2006)

No es extraño el uso de esta metodología en la revisión de información para los índices de precios y costos: El Instituto Nacional de Estadísticas chileno –INE- lo referencia como uno de los métodos de revisión del IPC –Índice de Precios al consumidor- aplicándolo a los mencionados relativos de precios (precio actual / precio anterior) (Instituto Nacional de Estadisticas, Febrero de 2009).

Grafica 1 . Representaciòn grafica del analisis de Tukey



Fuente: Grafico propio

La grafica 1 representa: Los atípicos a revisar corresponden tanto al 5% superior e inferior de la base, como aquellos que superan los límites superiores e inferiores. La media truncada de la parte superior de los datos se reconoce como la media truncada superior (S), en tanto que su contraparte responde como media truncada inferior (I)

2.2. METODO DE DISTANCIA MEDIA (UNITED NATIONS, 2009)

Este método se basa en la determinación del primer, segundo y tercer cuartil de la base. Dado que el análisis tiene en cuenta los cuartiles, la presencia de un valor extremo no afecta sus resultados en la magnitud observada al usar la media (OECD; FMI; OIT; NU; BM, 2006). La determinación del punto de corte a partir del cual se establece el valor atípico viene dado por:

Límite superior: Mediana + (C * Distancia media)
Límite inferior: Mediana - (C * Distancia media)

En donde C permite predeterminar o ajustar la proporción de registros a señalar como atípicos, así como definir previamente el número de veces la desviación estándar a tipificar:

Tabla 1. Valores de C y proporción de registros marcados. Distancia media

Valores de C y proporción de registros marcados

C	Como múltiplo de la desviación estándar	Proporción esperada de registros marcados
1	0,68	50%
2	1,37	17%
3	2,07	4%
4	2,75	0,70%
6	4	0,14%

Fuente: Practical Guide To Producing CPI (Página 192)

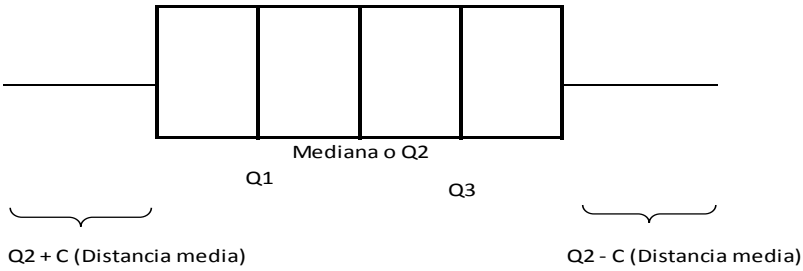
Traducción propia

Es importante considerar que aunque C es un parámetro previamente seleccionado por el investigador, puede tratarse o no un numero entero. Regularmente se utiliza el entero 2, el valor que será utilizado en el presente documento

Distancia media

Otro componente en la fórmula de determinación de atípicos es la distancia media, definida como un promedio simple que relaciona la distancia observada entre: a) el tercer cuartil y la mediana y b) la mediana y el primer cuartil.

Grafica 2. Representación gráfica del análisis de Distancia Media



Fuente: Grafico propio

En donde los valores sujetos de revisión corresponden a aquellos que superen la mediana (Q2) más o menos, el producto de C por la distancia media o el promedio entre la diferencia del tercer y segundo cuartil y éste con el primero.

2.3. GRAFICA DE CAJAS Y BIGOTES

La grafica de cajas y bigotes es construida, entre otros, a partir de la determinación de un valor crítico que define los atípicos de la base. Parte del análisis de cuartiles en su construcción y describe como atípico los valores que sobrepasen:

Límite superior: tercer cuartil + 1,5 (Rango intercuartilico)

Límite inferior: primer cuartil – 1,5 (Rango intercuartilico)

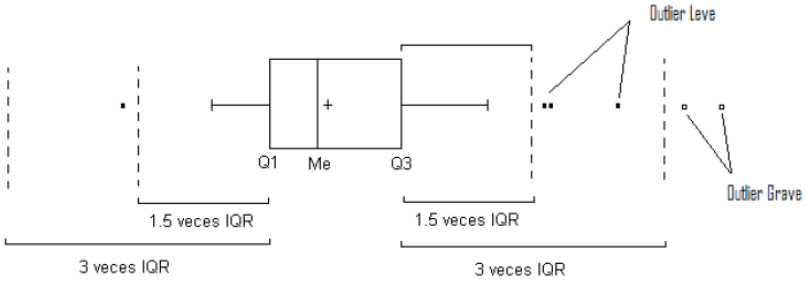
En donde el rango intercuartilico se define como la distancia entre el tercer y primer cuartil, por lo que incluye la mitad de todas las observaciones.

Dentro de la literatura es posible encontrar referencias a las caracterizaciones de los outliner: en caso en que éstos se establezcan al tener en cuenta el rango intercuartilico multiplicado por 1,5 se describen como errores, pero si se ubican tres veces el rango intercuartillico, se trata de errores gravísimos (Uribe, 2010):

Límite superior: tercer cuartil + 3 (Rango intercuartilico)

Límite inferior: primer cuartil – 3 (Rango intercuartilico)

Grafica 3. Representación gráfica: análisis de cajas y bigotes



Fuente: (Uribe; 2010)

La grafica muestra una caja que describe el punto en donde se encuentra la media de las observaciones (segundo cuartil), limitada por el primer y tercer cuartil. Por fuera de la caja de establecen las observaciones ubicadas con hasta 1,5 veces el rango intercuartilico y más allá, las ubicadas superando las tres veces el mismo rango. La línea ubicada por fuera de la caja es denominada el bigote (en plural describiendo tanto la parte inferior como superior de los datos).

2.4. PRUEBA DE DIXON

Esta prueba requiere el ordenamiento de los registros de mayor a menor y la revisión de los datos atípicos calculando la distancia del registro en revisión, respecto de: el siguiente valor considerado en la distribución, y el mínimo valor consignado:

a. Distancia del registro en revisión – el dato inmediatamente anterior
b. Distancia del registro en revisión - el mínimo dato de la base

A continuación se determina un estadístico de prueba calculado como la relación entre a y b: relación entre la distancia entre el atípico y el dato inmediatamente anterior y la distancia entre el atípico y el mínimo de la base. El resultado se compara con un valor critico obtenido en una tabla: en caso en que el estadístico resulte superior al valor crítico se entiende que el registro corresponde a un atípico; en caso en que el valor calculado sea inferior al descrito en la tabla, el registro es considerado normal (no atípico)

El valor critico depende de dos consideraciones: el nivel de confianza o el riesgo de calificar un registro como atípico cuando no lo es (Uribe, 2010). Para el presente trabajo el riesgo de no calificar como atípicos registros que si los son, es del 0,01%

3. TIPO DE ESTUDIO

El estudio realizado corresponde al tipo descriptivo, dado que se evaluarán los mecanismos para encontrar los valores atípicos en las base de datos que permiten calcular un índice de costos, (Hernàndez, Fernàndez, & Baptista, 1997); sin embargo no permite establecer las causas de dichos atípicos.

4. UNIDAD DE ANALISIS

Los datos requeridos para adelantar el trabajo refieren a las bases de precios que permitieron generar el cálculo de un índice de costos para el primer y segundo semestre de 2014, y primer semestre de 2015. Se utilizan las bases en dos momentos: la que describe la información analizada por el personal logístico (Equipo encargado del análisis de la información recolectada en cada ciudad y cuya base se nombrara como la base sin revisión, dado que aún no ha sido verificada desde temática: -equipo encargado de asegurar el cálculo, producción y difusión del índice-), y en segunda instancia la base **después** del análisis (dicha base se nombrara como la base resultado o grupo control)

El mecanismo de control a las metodológicas implica contrastar sus resultados con los obtenidos al revisar la totalidad de la información, estos resultados se encuentran en la base resultado de las revisiones de información.

Toda la información se contiene en archivos en Excel, con aproximadamente 4.000 registros por periodo de proceso (cada semestre). En dichas bases se incluye información del precio actual y anterior por artículo recolectado y se originan en la recolección de datos realizada por el personal a cargo de dichas tareas en 23 ciudades del país y para aproximadamente 560 fuentes diferentes (es posible que una sola fuente reporte varios precios en diferentes artículos de la canasta tomada por el índice).

El índice referido describe la variación promedio semestral de una canasta representativa de los bienes y servicios que requiere una institución de educación superior para dar cumplimiento a su objeto social. Requiere información de sueldos y salarios de personal docente, administrativo y general, así como los precios de diversos elementos, entre los que se destacan los servicios públicos, compra de software, libros, material de laboratorio, asociados a bienestar universitario, servicios de mantenimiento, reparación, vigilancia, aseo y publicidad entre otros.

Los precios recolectados afectan tanto a instituciones calificadas como universidades, instituciones universitarias, tecnológicas y técnicas de carácter tanto público como privado. (Las variaciones en los sueldos y salarios son recolectados teniendo en cuenta las diferencias entre estas ocho caracterizaciones descritas, que finalmente corresponden a los dominios de publicación del índice). La información calculada es utilizada por organismos del gobierno, Ministerio de Educación, instituciones de educación superior y entidades financieras, entre otros.

Los precios recopilados son analizados de manera tal que sea posible determinar si la variación observada en cada semestre corresponde a variaciones puras de precio (no determinadas por cambios en la calidad de los bienes y servicios). Aquellos que se refieran exclusivamente a cambios puros en el precio, permiten el cálculo del índice en su nivel elemental, que hace referencia a la agregación de datos para los artículos fuente a fuente, y posteriormente, la agregación de dichos elementos hasta llegar a un nivel de publicación. Los datos considerados como atípicos son revisados de manera

detallada de forma tal que se asegure la decisión sobre su posible exclusión de la base: la exclusión de un registro que describía un cambio relevante de precios, pero explicado únicamente por esa variable es un sesgo que debe evitarse, de la misma forma que la inclusión de variaciones explicadas por cambios de calidad.

Aunque la recolección y análisis local de la información se realiza en plataforma Oracle, es posible obtener las bases de datos finales en Excel de manera tal que es posible formular el trabajo descrito en el presente documento. El aplicativo permite determinar el perfil de usuario y hora en la que los diferentes involucrados toman decisiones (nivel local).

5. INSTRUMENTOS Y MATERIALES

Se requieren seis bases de datos, tres de ellas se refieren a los registros antes de revisión de atípicos para diferentes periodos (primer y segundo semestre de 2014 y primero de 2015). Cada base contiene aproximadamente 4.000 registros cada una, con información sobre los precios actuales y anteriores de los artículos recolectados Formato: Excel

Igualmente se cuenta con otro grupo de tres bases, referidas a los mismos semestres anteriormente denotados, pero que describen los valores marcados como atípicos cuando se ejecutó la revisión de toda la base, al verificar uno por uno cada registro. Estas bases son el resultado de la aplicación de la revisión que se adelanta actualmente y que requiere el análisis de la totalidad de la información.

6. PROCEDIMIENTO Y DISEÑO ESTADISTICO

El procedimiento partió con la aplicación de las tres diferente metodologías propuestas para la determinación de valores atípicos en las bases de datos antes de revisión.

Cada metodología arroja un número determinado de registros identificados como atípicos, esta información fue contrastada con los valores marcados como atípicos en las bases que contienen los resultados al hacer la revisión de la totalidad de la base.

Al finalizar este apartado del trabajo, fue posible determinar la metodología que: a) Obtuvo el mayor número de registros marcados como atípicos, y definidos como tal, cuando se revisó la totalidad de la base. b) El menor número de atípicos registrados como tal por la metodología propuesta, pero que no fueron identificados en la revisión de la totalidad de la base.

La conclusión sobre la mejor metodología aplicable para la detección de atípicos en las bases de relativos que permiten el cálculo de un índice de costos utiliza los criterios ya descritos, la seleccionada será aquella que presente el mayor número de registros marcados como atípicos, y definidos como tal por la base de control; y en segundo lugar, la metodología que registre el menor cantidad de atípicos pero que no coinciden con la revisión al 100%.

El desarrollo del trabajo utilizo Excel (Office de Microsoft Windows) y el complemento XLSTAT, versión 2015 con el fin de producir los resultados de la prueba de Dixon

Las bases de cálculo del índice de precios seleccionado, presentan algunas consideraciones importantes:

1. Únicamente se cuenta con tres procesamientos (bases), para realizar el análisis: primer y segundo semestre de 2014 y primer semestre de 2015. Aunque el indicador presenta resultados desde 1998, es desde el primer semestre de 2014 que se cuenta con información en la canasta rediseñada que permite la recolección en el momento actual.

   El mencionado rediseño afecto considerablemente los artículos seleccionados para hacer seguimiento a precios, por lo que no se pueden comparar los resultados posibles para bases de periodos anteriores y el presente.

   Dado que se trata de ubicar el histórico para intentar determinar cómo reacciona cada metodología de selección de atípicos, versus una base de control (la base que se revisa completamente y que permite el cálculo oficial del índice), no tiene mayor impacto la verificación de la metodologías para bases de semestres anteriores al primer semestre de 2014.

2. Se observa que la base correspondiente al primer semestre de 2014, es considerablemente inferior a la observada para los otros dos semestres disponibles: en total alcanza los 2.772 registros.

   La situación se soporta en el hecho que para el mencionado semestre se ubicaba una nueva canasta de recolección y es típicamente aceptado que en estas fases, el número de registros a recolectar sea sensiblemente inferior. Sin embargo el número recolectado fue suficiente para garantizar la representatividad suficiente en el cálculo de la variación promedio de precios. (En el segundo semestre de 2014 la base alcanzo los 4.204 registros, y en el primer semestre de 2015: 4.187)

3. Los precios recolectados permiten calcular la variación promedio para los salarios del personal asociado a la actividad en medición: (dichos artículos se definen como grupo 1); en tanto que los artículos designados como grupo 2, permiten determinar las variaciones promedio de los bienes y servicios necesarios para que dicha actividad sea posible: servicios públicos, arrendamientos, compras de bienes de oficina, seguros, dotaciones, servicios generales, transporte, etc.

   Dado el escaso número de semestres ocurridos, actualmente los precios de los salarios son recolectados para el primer y segundo semestre. Sin embargo, la metodología internacional sobre la materia (OECD; FMI; OIT; NU; BM, 2006), determina que la mejor practica implica el análisis de los precios recolectados en varios periodos y analizar la periodicidad de los cambios: en caso en que se evidencie que los precios se modifican cada año, es muy recomendable que la recolección de precios se realice en esa periodicidad, lo que implica que se disminuye la carga operativa para el instituto encargado y para las fuentes.

   Sin embargo, dado lo reciente del rediseño del índice en cuestión, este análisis no ha sido procedente, por lo que todos los precios son recolectados de acuerdo a la periodicidad de publicación: semestralmente, por lo que se evidencia que en segundo semestre existe sobreesfuerzo operativo, asociado a la recolección de los precios de los salarios, que típicamente tienen aumentos durante el primer semestre. La situación parece manifestarse también en los precios del grupo 2: bienes y servicios

4. Dentro de los protocolos de revisión aplicada a todos los registros recolectados, se incluye el uso de mallas de validación definidas directamente en las herramientas de recolección de la información (dispositivos móviles de captura), lo que permite hacer una primera validación de la información en el momento en que el recolector realiza su labor.

   Se pudo establecer que el grueso de las validaciones y consistencias aplicadas a los instrumentos de recolección fueron aplicada desde la instauración del rediseño. Lo anterior determina otra gran diferencia respecto de los registros

que se podrían recopilar de semestres anteriores a 2014, cuando el sistema de control y validación respondía a metodologías de revisión diferentes.

Sin embargo, y con el correr de los meses, el aplicativo de recolección y análisis de información tuvo ciertas modificaciones en sus mallas de validación, lo que permito afinar los procesos de los meses subsecuentes.

7. RESULTADOS OBTENIDOS

A continuación se describen los resultados obtenidos para las tres bases trabajadas:

7.1. PRIMER SEMESTRE DE 2014

El procedimiento seguido en la base del primer semestre implica el análisis de 2.772 registros en total. El uso de la metodología Turkey permitió generar un total de 404 registros para revisión, en tanto que el uso de las distancias medias determinó un total de 365, el uso de la metodología de cajas y bigotes ubicó un total de 164 registros a revisión y finalmente la aplicación de la prueba de Dixon determino 139. (Ver tabla 1)

Es muy interesante revisar la proporción de registros que tuvieron cambios después de analizar la totalidad de la base (51 registros): 1,84%. La caracterización de estos registros permite determinar que los cambios asociados se generaron por la aplicación incorrecta de novedades de recolección.

Las novedades de recolección se definen como el mecanismo técnico que permite afrontar cambios en las calidades de la información recolectada en la oficina de precios del DANE. Son varias las posibilidades a las que se enfrenta el equipo a cargo de la recolección de precios: que el bien o servicio ya no se ofrezca permanente o temporalmente en el mercado; que el bien o servicio tenga una variedad completamente comparable (un sustituto perfecto), en términos de la calidad; o que se haya un artículo completamente nuevo, razón por la cual, no existirá un precio anterior para calcular la variación de precios.

En el primer semestre de 2014, se observó un interesante número de observaciones que tenían asignada la novedad IN (Insumo nuevo), una situación apenas razonable dado que se trataba el primer semestre en donde se aplicaba la nueva canasta de recolección.

Para los 51 cambios observados en la base de control, 32 registros modificados por el nivel central, hacían referencia a la incorrecta adopción de la novedad IN, dado que para dichos registros el momento estadístico si posibilito encontrar el precio anterior y calcular la variación de precios. (El cambio realizado en el nivel central consistió en levantar la novedad IN y permitir que el registro ingresara en el cálculo oficial del índice, habilitando el precio anterior y generando la variación o relativo de precios).

Tabla 2. Comparativo resultado del primer semestre de 2014

Metodologia	1	2	3	4	5*
1	404				13
2		365			16
3			164		9
4				139	8
5*					51

**El resultado descrito en la columna 5 no hace referencia a una metodología de ubicación de atípicos sino de los resultados del número de registros que tuvieron modificación cierta, después de realizar la revisión de la totalidad de registros de la base (grupo control).*

Para los 19 registros restantes, se observó la necesidad de verificar el precio (supervisión y revisita a las fuentes), lo que finalmente concluyo en la ubicación de errores de toma de precios, que fueron subsanados con cambios de información y recalculo del relativo de precios.

Vale la pena denotar que el método que permitió ubicar el mayor número de registros en común versus el mecanismo de control (revisión de 51 registros en toda la base), fue el de distancias medias (16 coincidentes en total); en tanto que en segunda instancia se ubica la metodología Turkey. (13 registros), seguido de la metodología de cajas y bigotes que coincidió con 9 registros en total, finalizando con ocho los ocho registros definidos por medio del método de Dixon

La conclusión posible para este punto de la revisión implica que la metodología con mejor comportamiento (distancias medias), solo pudo detectar el 31% de los registros que ameritaban verificación y análisis prioritario.

7.2. SEGUNDO SEMESTRE DE 2014

La base observada para el segundo semestre de 2014 asciende a 4.204 registros en total, un fuerte incremento asociado a la madurez de la recolección en lo referente a aspectos de sensibilización y ajuste de directorios. Otro cambio importante observado en este semestre, tiene que ver con la afinación y mejoramiento de las mallas de validación y consistencia aplicadas al análisis básico de los datos capturados desde el dispositivo móvil de captura

En este punto se hace importante describir las características generales de la base: de un total de 4.204 registros, 3.455 tienen relativo el 1, lo que quiere decir que no existió variación de precios (el resultado se obtiene como un relativo de precios: precio anterior/ precio actual). La razón que soporta el comportamiento se encuentra en que como se mencionó anteriormente, la metodología del indicador obliga la recolección de todos los artículos de la canasta, de manera semestral, sin embargo, es claro que los

salarios, por ejemplo tienen concentrada su variación en el primer semestre (no en el segundo), y que lo propio también ocurre con ciertos bienes y servicios del grupo del mismo nombre (el asunto referido a los incrementos anuales en los contratos de seguridad, aseo y cafetería y/o restaurante son un claro ejemplo).

Los resultados obtenidos al determinar los atípicos de acuerdo a cada metodología se describen en la tabla 3.

Tabla 3 . Comparativo resultado para el segundo semestre de 2014

Metodologia	1	2	3	4	5*
1	420				
2		237			
3			101		
4				197	
5*					0

**El resultado descrito en la columna 5 no hace referencia a una metodología de ubicación de atípicos sino de los resultados del número de registros que tuvieron modificación cierta, después de realizar la revisión de la totalidad de registros de la base (grupo control).*

Es interesante observar el número de registros que fueron modificados por el nivel central para el segundo semestre de 2014. A pesar del incremento en el número de registros, no se ubicó ninguno que tuviera cambio en la información, cubierta la etapa de análisis para la totalidad de la base. De nuevo y como se observó para el primer semestre, el número de atípicos distinguidos aplicando la metodología de cajas y bigotes es menor al resto, en este caso solo 101 registros, seguido en su orden por la prueba de Dixon, (197 registros), la metodología de distancia media y de Turkey.

Sin embargo, conviene aquí describir el comportamiento observado en la base perteneciente al segundo semestre de 2014, y que finalmente llevo a afectar la metodología de cálculo aplicada en el ejercicio de distancias medias y cajas y bigotes:

Al intentar generar los valores para el primer y segundo cuartil (mediana), por ejemplo, se encontró que el resultado era igual a 1. Dada la situación, se optó por retirar la totalidad de registros cuyo relativo fuera igual a la unidad (sin variación de precios). Después de realizar dicho tratamiento, los valores del primer, segundo y tercer cuartil permitieron la aplicación de las metodologías propuestas, generando los resultados presentados.

Para este semestre en particular, se observa que realmente, no era necesaria la revisión de la totalidad de la base, ya que ninguno de los registros requirió cambio y fue incluido dentro del cálculo oficial del índice, por lo que la designación de la mejor metodología apela al criterio del menor número de registros por revisar, dado lo anterior, la escogida en este semestre es la cajas y bigotes, seguida de la prueba de Dixon, la de distancia media, y dejando en cuarto lugar la de Tukey. Hasta este punto no es posible ser partidario de una metodología, dado los resultados dispares logrados para las bases recolectadas en 2014.

7.3. PRIMER SEMESTRE DE 2015

Para la base del primer semestre de 2015, se presentaron resultados aún más interesantes, en este periodo la información debe mostrar incrementos de precios relacionados con el ajuste anual típico de este periodo del año, igualmente se cuenta con una metodología de recolección y validación en dispositivo más depurada, lo que dio por resultado una base total de cálculo de 4.187 registros, de los cuales finalmente fueron modificados un total de 11.

Tabla 4 . Comparativo resultado para el primer semestre de 2015

Metodologia	1	2	3	4	5*
1	518				11
2		796			11
3			558		11
4				243	6
5*					11

**El resultado descrito en la columna 4 no hace referencia a una metodología de ubicación de atípicos sino de los resultados del número de registros que tuvieron modificación cierta, después de realizar la revisión de la totalidad de registros de la base (grupo control).*

Contrario a lo ocurrido en el resto de los semestres, la totalidad de los registros que fueron ajustados en el grupo de control fueron percibidos también por tres de las cuatro metodologías propuestas.

En este caso, es la metodología de Dixon la que genero el menor número de registros a revisar, seguido de la de Tukey, cajas y bigotes y finalmente la de distancias medias.

De nuevo los resultados observados de acuerdo a los criterios de análisis de las metodologías difieren en este semestre, en relación a lo observado en las dos bases anteriores.

Sin embargo, se resalta como en los tres periodos analizados se han presentado situaciones diferenciales que soportarían los resultados obtenidos: se ha establecido como un primer periodo de recolección contiene menos registros, y un proceso de asentamiento de la canasta seleccionada; la segunda base, por su parte, demuestra el fortalecimiento de los mecanismos de análisis primarios que determinaron que la base tuviese robustez desde el punto de vista de análisis de atípicos (ninguno de ellos tuvo que ser modificado y se aceptó como un relativo que describía la realidad de la variación de precios), y finalmente se ubica una tercera base con un número importante de registros, mecanismos de análisis muy robustos y variaciones de precios cargados a las derecha de la distribución (en general se denotan para ese periodo, incrementos de precios).

Se espera que en semestres posteriores, se pueda contar con bases que mantengan las características denotadas para el primer semestre de este año, lo que equilibraría las condiciones de recolección, aunque se sugiere que para los segundos semestres se verifique la periodicidad de los cambios de precios, de forma tal que el desgaste operativo sea menor y la base final sujeta a análisis contenga relativos iguales a uno, en justas proporciones. Es posible que después de tener la quinta base (primer semestre de 2016), el análisis de la metodología determinación de atípicos ofrezca resultados más estables en el tiempo.

De igual forma, la frontera de tiempo propuesta (un año a partir de la fecha) es el tiempo requerido para que se implemente la sistematización de cualquier proceso de análisis, previa parametrizaciòn de los criterios que definirán los atípicos a revisar prioritariamente.

CONCLUSIONES

Después de finalizar el presente trabajo de investigación es posible generar algunas conclusiones y recomendaciones al proceso de análisis de información para el índice de precios estudiado.

1. Es necesario aplicar un método estadístico de revisión a la información, ya que es evidente el desgaste que debe ejecutar el equipo temático a cargo de la investigación al tener que revisar la totalidad de la base cuando la proporción máxima de cambios que permite esa tarea es de menos del 2% (51 registros de 2.772 para el primer semestre de 2014).

2. A pesar de contar con tres bases de cálculo diferentes, cada una de ellas representa una situación particular: el primer semestre de 2014 pareciese tratarse de un periodo de ajuste, generado como consecuencia de la actualización de la canasta aplicada en el rediseño del indicador: el número de registros presentes en la base era suficiente para representar la variación promedio de precios, pero es menor a la observada en el resto de periodos y la novedad asociada a los insumos nuevos tuvo que ser verificada desde el nivel central para procurar el ingreso de registros que contaban con el precio anterior.

   Respecto al segundo semestre de 2014, no hay comparación posible con el mismo periodo para otro año, sin embargo es evidente que la periodicidad de la recolección de precios debe ajustarse a su variabilidad, con el fin de evitar sobrecargar la recolección en semestres en donde no se ejecutan cambios. Lo anterior impactaría la base de revisión.

   Finalmente y respecto al primer semestre de 2015, se podría detectar cierta madurez en la toma de precios: los atípicos detectados en la variación total de la base, fueron ubicados a partir del uso de tres de las cuatro metodologías propuestas.

3. Los resultados observados no parecen ser consistentes con la toma de decisión en favor de una de las cuatro metodologías propuestas: durante el primer periodo, el mayor número de atípicos detectados ubica la metodología de distancias medias, como la más interesante (siempre priorizando que sea posible detectar los registros que realmente requieren la intervención del nivel central). En la segunda base sin embargo, y dado que no se realizaron modificaciones en la base control, la metodóloga de cajas y bigotes resulta ser la elegida, dado que presenta el menor

número de registros a verificar. Finalmente, en la tercera base, y dado que las cuatro metodologías permitieron identificar los registros que ciertamente requieren el análisis, la más interesante resulta ser la metodóloga de Turkey, ya que genera la menor cantidad de registros por verificar, y el mayor número de registros coincidentes con el grupo control, superando la prueba de Dixon, que definió menos registros por revisar, pero que seleccionó menos registros que realmente requerían revisión

4. Como conclusión general se recomendaría continuar con el proceso de análisis de las bases posteriores del indicador, con el fin de afianzar la toma de decisión respeto a una en particular.

5. Se hace relevante que los encargados del indicador análisis las variaciones de precios y sus periodicidades con el fin de reducir la carga operativa que implica recolectar el 100% de la canasta en todos los semestres. La evidencia advierte que por lo menos un segmento muy importante de precios debe cambiar cada año.

6. De otro lado, y en la medida en que sea posible su implementación, se recomienda la revisión de métodos alternos que permitirán mejorar las conclusiones, tales como el uso de la regresión y la metodología de detección de atípicos de Mahalanobi. (mayor información incluida en el anexo A y B del presente documento)

ANEXO A

El método de regresión consiste en ajustar los datos para determinar una ecuación, de forma tal que se modele una relación entre variables. La forma de regresión más sencilla es la lineal y en el estudio de outliners se utiliza observando aquellos puntos que se encuentran más alejados de los resultados fijados por el modelo generado, para el caso de la regresión lineal, se estudiarían aquellas observaciones más alejadas a las generadas por la ecuación (línea)

La regresión lineal requiere que exista alguna relación entre las variables, de forma tal que se pueda establecer una variable dependiente (que se mueve en función del comportamiento de otras) y otra variable independiente:

X = parámetro autónomo + parámetro de la variable (Y) + e*

La variable dependiente X, se explica a partir de dos componentes: un parámetro autónomo, que no depende de una variable y de cómo se comporte la variable independiente (parámetro de la variable Y), agregando además un contenedor del error (e*). Por ejemplo, la tarifa de un servicio público (electricidad) depende de un costo fijo y del costo variable multiplicado por el número de kilovatios consumidos, en donde Y correspondería a dicho consumo (kv)

Típicamente la construcción de un modelo de regresión requiere establecer una relación o asociación clara entre variables, lo que dificulta su utilización en el análisis de outliner de relativos de precios para un índice, dado que la recolección de información únicamente incluye el comportamiento de los precios durante el tiempo (no existe otras variables a incluir), y no es adecuado asumir que los mismos mantengan una relación –asociación- típicamente positiva o negativa con el correr de los días, semanas o meses (dependiendo lo anterior de periodicidad de recolección de información).

La determinación apropiada de la relación entre variables es un paso primordial, dado que los puntos lejanos a la recta descrita en la ecuación, pueden deber dicho comportamiento a que efectivamente son valores atípicos, o a que la ecuación no describe de la manera más adecuada las relaciones entre variables, y simplemente sea un síntoma de que el modelo requiere revisión.

Posteriormente, se establece el mejor método posible para ajustar los datos a un modelo particular, el más conocido es el método de mínimos cuadrados, que consiste en minimizar la diferencia entre el valor calculado por la regresión y el dato realmente consignado en la base (la diferencia se potencializa al cuadrado), eligiéndose aquel modelo que minimiza la suma de cada diferencia calculada. (Uribe, 2010)

ANEXO B

El método Mahalanobis utiliza el análisis de distancias calificando un dato atípico como aquel que tiene la mayor distancia respecto de la nube de puntos (masa en donde se ubican la mayoría de los puntos); por el contrario un punto dentro de la masa no tiene distancia de diferencia y es calificado como un registro normal.

Se diferencia de la tradicional medición de distancias euclidianas porque tiene en cuenta la correlación entre variables. La distancia es calculada para cada registro, asignando a cada distancia un peso: aquellas distancias superiores (datos más atípicos), se valoran con un peso inferior a aquellas distancias inferiores (no atípicos), los pesos relativos son usados para calcular una regresión ponderada que busca minimizar el efecto de los valores extremos.

La distancia de Mahalanobis sigue una distribución chi-cuadrado y sus grados de libertad son iguales al número de variables incluidas en el cálculo. (Uribe, 2010)

REFERENCIAS

DANE. (2013). *Metodologia Indice de Precios al Consumidor.* Bogotà.

Hernàndez, Fernàndez, & Baptista. (1997). *Metodologìa de la Investigaciòn.* Bogotà: McGRAW - HILL.

Instituto Nacional de Estadististicas. (Febrero de 2009). *Manual Metodologico del ìndice de precios al consumidor nacional.* Santiago de Chile: http://www.ine.cl/canales/chile_estadistico/estadisticas_precios/ipc/metodologia/31_03_10/Manual%20Metodologico%20NIPC%20BASE%20ANUAL%202009.pdf.

OECD; FMI; OIT; NU; BM. (2006). *Manual del ìndice de precios al consumidor.* Washington: Departamento de Tecnòlogia y servicios generales.

Saìdi. (2005). *Detection of Outliers in the Canadian Consumer Price Index.* Ottawa: Comisión Estadística de las Naciones Unidas y Comisiòn Economica Europea.

UNITED NATIONS. (2009). *PRACTICAL GUIDE TO PRODUCING CONSUMER PRICE INDICES.* NEW YORK AND GENEVA: UNITED NATIONS.

Uribe, I. A. (2010). *Guia Metodològica para la Selecciòn de Tècnicas de Depuraciòn de Datos.* Medellìn: Universidad Nacional de Colombia.

Vena, P. (2014). *Detecciòn de datos atipicos para datos funcionales asimetricos.* Buenos Aires: Tesis de Licenciatura.