



*Asesorías y Tutorías para la Investigación Científica en la Educación Puig-Salabarría S.C.  
José María Pino Suárez 400-2 esq a Lerdo de Tejada. Toluca, Estado de México. 7223898475*

RFC: ATI120618V12

**Revista Dilemas Contemporáneos: Educación, Política y Valores.**

<http://www.dilemascontemporaneoseducacionpoliticayvalores.com/>

ISSN: 2007 – 7890.

**Año: IV. Número: 2. Artículo no.45 Período: Octubre, 2016 - Enero, 2017.**

**TÍTULO:** Diagnóstico de la calidad de los datos siguiendo el enfoque de la ISO 9001:2015.

**AUTORES:**

1. Ing. Liset Agüero Zardón.
2. Dr. José A. Vilalta Alonso.
3. Dra. Rosario Garza Ríos.
4. Ing. Idalianys Urquiola García.

**RESUMEN:** En Cuba, se han desarrollado algunas investigaciones que arrojan como principal resultado que en la mayoría de las empresas no existen evidencias de la realización de estudios sobre la calidad de los datos. Dada esta situación, en el año 2008 se presenta un doctorado que propone un procedimiento de diagnóstico. No obstante, un análisis detallado del mismo permitió detectar determinadas limitaciones, entre las que se destaca que no tiene en cuenta la nueva versión de la ISO 9001:2015 en cuanto a Enfoque a Riesgos. De ahí que se defina como objetivo general del estudio: Rediseñar el Procedimiento para Diagnosticar la Calidad de los Datos.

**PALABRAS CLAVES:** calidad de los datos, dimensiones de calidad de los datos, enfoque a riesgos.

**TITLE:** Diagnosis of data quality following the ISO 9001:2015 approach.

**AUTHORS:**

1. Eng. Liset Agüero Zardón.
2. Dr. José A. Vilalta Alonso.
3. Dra. Rosario Garza Ríos.
4. Eng. Idalianys Urquiola García.

**ABSTRACT:** There have been carrying out some researches which have as a result that there isn't any evidence of this study in most of the Cuban enterprises concerning the quality of data that are used in the decision-making process. Due to this, there is a doctorate thesis that proposes a procedure for the diagnostic of data quality in 2008. However, a detailed analysis made possible to recognize certain limitations such as that the new version ISO 9001:2015 concerning risk approaches is not taking into account. For this reason, the general objective of this study is defined as: Redesign the procedure to diagnose data quality.

**KEY WORDS:** Data quality, dimensions of data quality, risks approach.

**INTRODUCCIÓN.**

Los datos se emplean para apoyar la mayoría de las decisiones organizacionales, sean estas operacionales, a nivel de dirección o estratégicas (Redman, 2004). Cada fuente de datos tiene una variedad de valores o campos que faltan, entre otros errores, que son el centro del problema que conduce a una carencia total de la calidad de los datos (Haro, 2007).

Los datos inexactos o contrarios pueden obstaculizar la capacidad de la organización de entender sus problemas de negocio actuales y del futuro (Calero, 2007). Esto conduce a decisiones pobres que pueden causar una serie de resultados negativos, incluyendo pérdida de beneficios, operaciones retrasadas, descontento de los clientes y mucho más (Ortuño, 2005).

La mala calidad de los datos influye de manera muy significativa y profunda en la efectividad y eficiencia de las organizaciones, así como en todo el negocio, llevando en algunos casos a pérdidas multimillonarias (Bianchi, 2009). Es por ello la importancia que actualmente se atribuye a la toma de decisiones basada en datos confiables, es decir, que cumplan con los requisitos de calidad establecidos para su uso.

El pensamiento basado en riesgos permite a una organización determinar los factores que podrían causar que sus procesos y su sistema de gestión de la calidad se desvíen de los resultados planificados, para poner en marcha controles preventivos, para minimizar los efectos negativos, y maximizar el uso de las oportunidades a medida que surjan (Véase el capítulo A-4, Norma ISO 9001:2015). Abordar tanto los riesgos como las oportunidades establecen una base para aumentar la eficacia del sistema de gestión de la calidad, alcanzar mejores resultados, y prevenir los efectos negativos.

## **DESARROLLO.**

### **Rediseño del procedimiento para el diagnóstico de la calidad de los datos.**

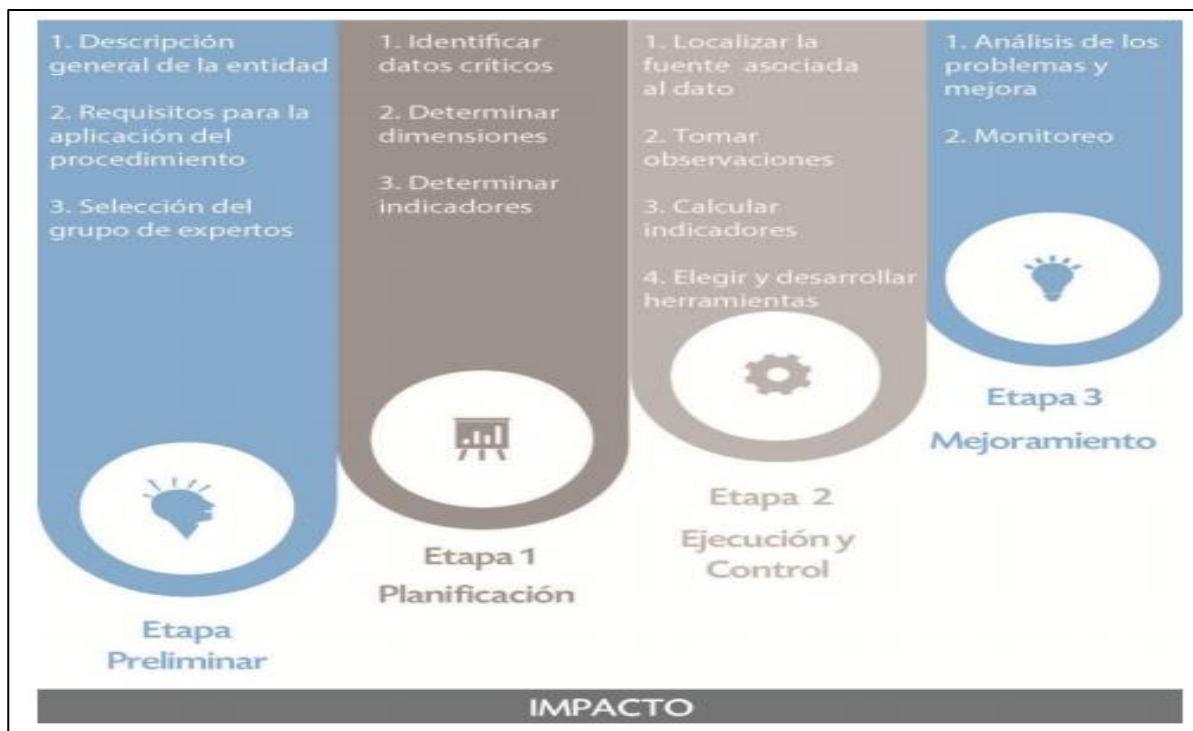
El procedimiento que se toma como base de esta investigación fue propuesto originalmente por (Vilalta, 2008) como resultado de su tesis de doctorado. Por este motivo, se presentará una breve descripción del mismo (incluyendo el rediseño propuesto), el cual puede ser consultado en:

<http://hdl.handle.net/123456789/1411>

Este procedimiento está estructurado en tres etapas fundamentales, en las que se especifican los objetivos individuales y los principales resultados a obtener; además de una etapa preliminar en la que se define, entre otros elementos, el grupo de trabajo que se conformará para apoyar la aplicación.

Tras un análisis detallado del procedimiento, se determinó rediseñar aspectos considerados como limitantes del mismo, en cuanto a: definir los datos críticos como aquellos que consideran de riesgo, enfatizándose previamente en la perspectiva que se sigue con este enfoque, así también como la implantación de un plan de prevención de riesgos que sea de utilidad para la mejora del funcionamiento de la empresa, y medir el impacto que pueda representar para la empresa el uso de datos de mala calidad de manera más cuantificable.

Las etapas del procedimiento se muestran en la siguiente figura:



**Figura 1:** Procedimiento para el diagnóstico de la calidad de los datos.

**Fuente:** Elaboración propia a partir de Vilalta, 2008.

Cada una de estas etapas se describe detalladamente a continuación.

### **Etapa preliminar.**

Esta etapa tiene como objetivo principal realizar una descripción general de la entidad objeto de estudio, así como la aclaración de ciertas condiciones que son pilares esenciales, relacionados con la responsabilidad, el compromiso de la dirección, la capacitación del personal involucrado

con la aplicación inicial del procedimiento, y posteriormente, para todo el personal implicado en el manejo de los datos en la empresa.

Es necesario y de carácter obligatorio, para la aplicación efectiva del procedimiento, la conformación de un grupo de trabajo, en el cual sea considerado el personal como experto en el tema de la investigación. El mismo debe apoyar la implementación del procedimiento aportando sus criterios en varios aspectos.

### **Etapas 1: Planificación.**

Los objetivos específicos de esta etapa son la detección de los datos críticos, la determinación de las principales dimensiones de calidad de los datos, así como la definición de los indicadores que van a servir para medir estos criterios. El responsable de esta etapa será algún consultor externo o representante de la entidad. Esta etapa incluye las siguientes actividades:

➤ **Identificar datos críticos.**

Se tienen en cuenta, para el diagnóstico, aquellos datos considerados como críticos; es decir, los datos más importantes o que mayor impacto tengan sobre las decisiones que se toman según un uso determinado del propio dato. También se puede definir como dato crítico a aquel dato que presente mayor cantidad de errores o problemas y con más frecuencia que el resto. Existe un concepto que encierra las dos definiciones de dato crítico mencionadas anteriormente, que es el dato de mayor riesgo.

Se propone utilizar como forma de identificación, ordenamiento y selección de los datos críticos, la información adquirida de las entidades y/o miembros de grupo de trabajo, pero que a la vez presente un enfoque a riesgos. Esta propuesta se basa en la fundamentación de la vigente Norma ISO 9001:2015 en cuanto a enfoque a riesgos, donde en el momento en que una organización planifica su sistema de gestión de la calidad debe considerar la determinación de los riesgos y

oportunidades, con el propósito de prevenir efectos no deseados, lograr la mejora del sistema de gestión, y asegurar que el mismo logre los resultados pronosticados.

En las organizaciones existen, por lo general, una gran variedad de tipos de datos, por lo que esta actividad se realizará comenzando por un listado con los posibles datos críticos a partir de su consideración como riesgo. La medición del riesgo se propone definirla por tres variables: frecuencia, impacto y eficacia de los controles (las variables a analizar pueden variar según los intereses de la empresa que aplique el procedimiento), tomando las siguientes escalas:

- Frecuencia de los eventos:

- Alta (el evento ocurre una vez al trimestre, este nivel tendrá una calificación de 5),
- Media (el evento ocurre una vez al semestre, este nivel tendrá una calificación de 3),
- Baja (el evento ocurre con una periodicidad superior al semestre, este nivel tendrá una calificación de 1).

- Impacto de la materialización de los eventos:

El impacto del evento se califica de acuerdo a las siguientes categorías, asignándosele uno de los siguientes valores:

**Tabla 1:** Escala para medir el impacto de la materialización de los eventos.

<b>Categoría</b>	<b>Valor</b>
Catastrófico	5
Moderado	3
Leve	1

**Fuente:** Elaboración propia.

Para asignar la categoría, se emplea la Matriz de Clasificación de los Riesgos que se presenta continuación:

**Tabla 2:** Matriz de Clasificación de los Riesgos.

	<b>Catastrófico</b>	<b>Moderado</b>	<b>Leve</b>
Objetivos/Proyectos/ Procesos	El evento afectó el cronograma establecido para uno o varios de los objetivos primarios, generando una ampliación en las fechas para el cumplimiento de los mismos.	El evento afectó el cronograma establecido para uno o varios de los objetivos secundarios, generando una ampliación en las fechas para el cumplimiento de los mismos.	El evento tuvo un efecto menor en el alcance de los objetivos.
Funcionalidad del dato	El evento afecta totalmente la funcionalidad del dato.	El evento afecta parcialmente la funcionalidad del dato.	El evento afecta poco o nada la funcionalidad del dato.
Impacto económico	Evento que tiene un elevado impacto económico.	Evento que tiene un impacto económico medio.	Evento que tiene un menor impacto económico.
Toma de decisiones	Evento que afecta la toma de decisiones.	Evento que afecta en parte la toma de decisiones.	Evento que no afecta o prácticamente no afecta la toma de decisiones.
Dependencia de otros datos	Alto nivel de dependencia de otros datos, del dato afectado por este evento.	Bajo nivel de dependencia de otros datos, del dato afectado por este evento.	Muy poco o ningún nivel de dependencia de otros datos, del dato afectado por este evento.
Quejas y reclamaciones	Se impusieron demandas en contra de la entidad por fallas en la prestación del servicio, con indemnizaciones por parte de la entidad.	Se recibieron múltiples quejas por mala gestión o faltas en la prestación del servicio de la entidad.	Se recibieron quejas por demoras en la prestación del servicio de la entidad fácilmente solucionables.

**Fuente:** Elaboración propia a partir de Vilalta, 2008.

Luego de conocer la ubicación de cada tipo de dato en la clasificación del riesgo por cada impacto o consecuencia, se realiza el cálculo de la función de ponderación, cuya expresión es:

$$F(d) = C * P$$

Donde:

F(d): es la función de ponderación del dato.

C: es la cantidad de veces que cada tipo de dato se ubica en la clasificación del riesgo.

P: es el peso en cada clasificación del riesgo.

En caso de existir empate en el resultado de la función de ponderación para un mismo tipo de dato, en diferentes clasificaciones del impacto, la valoración se debe hacer por el mayor inmediato, exceptuando aquellos casos donde el resultado de la función de ponderación en la clasificación de Catastrófico supere el existente en los empates de menor clasificación del riesgo.

- Detección de los eventos a través de la eficacia de los controles:

El factor Detección se traduce como la eficacia de los controles existentes en la empresa para detectar el riesgo antes de que este ocurra. El mismo se califica de acuerdo a los siguientes criterios, en base a una escala de 5 puntos de acuerdo a la tabla que se muestra, de modo que la mayor puntuación corresponda al riesgo que menos probabilidad presenta de ser detectado, debido a que no se aplican medidas para la prevención y el control, en cambio la menor puntuación se refiere a los riesgos que más probabilidades presentan de ser revelados, debido a que los controles realizados son los adecuados.

**Tabla 3:** Escala para medir la detección de los controles existentes.

Valor	Nivel de detección	Criterio
1	Alto	Los controles existentes del proceso son eficientes por lo que generalmente siempre se detecta el riesgo con antelación suficiente.
3	Moderado	Los controles existentes del proceso son mejorables, por lo que la probabilidad de que los riesgos sean detectados es media.
5	Bajo	Los controles existentes del proceso son deficientes y/o insuficientes, por lo que existen pocas posibilidades de que sea detectado con antelación suficiente.

**Fuente:** Elaboración propia.

Tomando los valores de los aspectos analizados para la definición de los datos críticos, se calcula el Índice de Riesgo del Dato (IRD), que no es más que el resultado de la multiplicación de los



tres valores. Este índice permitirá establecer un orden de prioridad para la selección de los datos que serán diagnosticados.

Posteriormente, a los resultados que arroje el índice se les otorgará un nivel de prioridad para su estudio. Este nivel será primeramente INACEPTABLE, seguido por IMPORTANTE, MODERADO, TOLERABLE y por último ACEPTABLE.

➤ Determinar dimensiones.

Luego de obtener los datos críticos definitivos a diagnosticar, las dimensiones de éstos son los atributos que definen su calidad, y las que se utilicen deben estar relacionadas con las características de los datos en uso.

El personal encargado de aplicar el procedimiento deberá confeccionar un listado con las dimensiones que serán utilizadas en la evaluación de la calidad de los datos. Entre ellas deben estar la exactitud, integridad, y actualidad (a menos que hayan razones que justifiquen otra cosa) y que será completada con otras (tales como la coherencia, la consistencia, la puntualidad, etc.), que a su juicio, puedan ser también necesarias, teniendo en cuenta las características de los datos del proceso y los requisitos de calidad de datos de los usuarios.

La lista finalizada será entregada a los miembros del grupo de trabajo y se procederá a aplicar el método Delphy por rondas, para depurarla y ordenarla según el grado de importancia de las dimensiones, para lo cual se propone que no existan ligas o empates para no reducir la capacidad de ordenamiento.

➤ Determinar indicadores.

Para comenzar, se propone la creación de un indicador que refleje el porcentaje total de datos correctos, calculado como la fracción de datos libres de errores del total de datos analizados.

De forma general, se deben definir indicadores que reflejen la fracción de datos que cumplen con la definición dada para cada dimensión analizada. Los indicadores a utilizar, asociados a las dimensiones exactitud, integridad y actualidad, son:

- Exactitud: Porcentaje de valores que están correctos cuando se comparan los valores actuales con los de referencia.
- Integridad: Porcentaje de datos que poseen todos sus elementos cuando se comparan los valores actuales con los de referencia.
- Actualidad: Porcentaje de datos que están actualizados cuando se comparan los valores actuales con los de referencia.

Además, se podría calcular un coeficiente ponderado, al que se le llamaría Índice Ponderado de Calidad del Dato (IPCD), dado que se tienen las ponderaciones de las dimensiones dadas por los expertos. Este índice permitirá la realización de una evaluación del nivel de calidad de los datos más integral, en el que se incluye la situación de cada una de las dimensiones consideradas.

La expresión para calcular este índice a partir del orden orientado por los expertos, en caso de existir concordancia entre ellos es:

$$IPCD = \sum Di * Wi$$

Donde:

Di: Valor, en porcentaje, de la dimensión i (valor de los indicadores calculados).

Wi: Peso de la dimensión i obtenido a partir del ordenamiento dado por los expertos.

Se podrán definir indicadores asociados a los tipos de datos, la gravedad de los errores, y cualquier otro que se considere necesario.

**Etapas 2: Ejecución y control.**

Los objetivos de esta segunda etapa son medir la calidad actual de los datos críticos definidos anteriormente, identificar los principales problemas de calidad que afectan al dato, y llevar un control del comportamiento de los indicadores establecidos para así promover de forma preventiva la calidad del dato. Como resultados se obtendrán el valor real de los indicadores asociados a cada dimensión definida, un resumen de problemas detectados, las causas que los provocan, así como la evaluación de los planes de medidas tomados con anterioridad, una vez que el procedimiento se esté aplicando sistemáticamente.

El responsable de la etapa será el consultor externo, especialista de calidad u otro trabajador capacitado. Esta etapa incluye las siguientes actividades:

➤ Localizar la fuente asociada al dato.

La fuente de un dato es la base o registro donde se encuentran los tipos de datos que van a ser analizados, que debe ser aquella que se usa en la toma de decisiones. Para cada tipo de dato debe localizarse aquella fuente que sirva de referencia de calidad, que pudiera ser la misma para todos o no. La presencia de un error en un dato estará dada por la diferencia en los valores que toma el dato en dos fuentes diferentes o la diferencia entre los valores que toma en un registro y los que realmente posee la entidad que ha sido caracterizada con esos datos.

➤ Tomar observaciones.

Se debe comenzar con el cálculo del tamaño de muestra, con la cual se verificará la existencia de desviaciones en las dimensiones propuestas, y a partir de la cual se calcularán los indicadores que reflejarán el estado actual de la calidad de los datos.

Para calcular el tamaño de muestra se debe partir de analizar que lo que se está valorando es un porcentaje, y que además, se quiere precisar un error en la estimación, por lo que es común hacer uso de un proceso iterativo para lograrlo.

➤ Calcular indicadores.

Los datos analizados deben localizarlos en la fuente de uso y verificar la calidad del tipo de dato que se esté analizando, comparándolo con la fuente patrón de calidad, anotando detalles en una tabla como la siguiente:

**Tabla 4:** Resumen de problemas detectados.

<b>Dato</b>	<b>Descripción del error</b>	<b>Dimensión afectada</b>	<b>Clasificación</b>
Dato 1	...	...	...
Dato 2	...	...	...
Dato n	...	...	...

**Fuente:** Vilalta, 2008.

La tabla anterior permite determinar el porcentaje de datos correctos; es decir, el porcentaje del tipo de dato analizado que cumple con cada una de las dimensiones a partir de los indicadores definidos. También se puede conocer la proporción del total de errores que son graves, de gravedad media y menos graves, aunque para este criterio de gravedad es relevante la participación del grupo de trabajo capacitado que debe asignarle una gravedad determinada a cada caso.

Para calcular los indicadores se procede como se explicó en la etapa 1.

➤ Elegir y desarrollar herramientas.

En una primera aplicación del procedimiento no se tendrá disponible información previa, por lo que no existirá la posibilidad de realizar comparaciones entre tipos de datos, dimensiones, períodos precedentes, etc. Al carecerse de una base comparativa necesaria como para analizar el desarrollo de los indicadores asociados a las dimensiones de calidad del dato, no se podrá arribar a ninguna conclusión completa acerca de la calidad de los datos. Por lo tanto, este análisis será más rico y alcanzará su plenitud, en la medida en que se vayan haciendo aplicaciones sucesivas del procedimiento.

En esta actividad se debe tener en cuenta el análisis de los indicadores asociados a las dimensiones, el comportamiento de los errores entre los diferentes tipos de datos, y la gravedad de estos. Se pueden utilizar gráficos pastel, diagramas de barras o el diagrama Pareto, para resaltar el comportamiento de indicadores. Para situaciones futuras se puede crear una base de datos histórica con los principales problemas para emplearla doctamente en una correcta toma de decisiones.

### **Etapas 3: Mejoramiento.**

Esta etapa tiene como objetivos determinar la importancia relativa de cada problema de calidad del dato, el lugar dentro del proceso donde se origina, y determinar las causas que provocan los problemas de calidad, proporcionando vías para la mejora y monitoreando su comportamiento. Otro objetivo será evaluar la efectividad de los planes de medidas anteriores, una vez que se esté aplicando el procedimiento sistemáticamente. Se espera obtener como resultados la importancia de los problemas, la ubicación del área o parte del proceso donde se producen estos problemas, las causas principales de los problemas de calidad, los planes de mejora, y una evaluación de los anteriores, cuando proceda. El responsable será el consultor externo, especialista de calidad o trabajador capacitado designado por la dirección. Las actividades que la integran son las siguientes:

- Análisis de los problemas y mejora.

A partir de todo el trabajo realizado se obtendrá como resultado un grupo de problemas con la calidad de los datos, relacionados con las dimensiones, los tipos de datos y los errores que con mayor frecuencia están presentes, así como el nivel de gravedad de los mismos, lo que brinda un determinado orden de prioridad para enfrentar estos problemas.

Para la eficaz determinación de las posibles causas asociadas a los problemas de calidad del dato es importante aterrizar en las áreas del proceso, donde se generan los errores en los datos. Se deben definir las barreras generales de prevención contra errores en los datos, las cuales pueden estar constituidas por las personas, la tecnología informática, los procesos de trabajo, las normas y regulaciones, y el ambiente de trabajo; para posteriormente buscar en el estudio específico que se haga, las causas que afectan cada barrera.

Tras identificar las causas que generan los errores en los datos, se deben trazar planes de mejora para los indicadores y eliminar dichas causas. A pesar que la mejora de la calidad de los datos tiene sus peculiaridades en cada uno de los procesos donde ésta se analice, algunos elementos a tener en cuenta pueden ser el diseño de una base de datos, duplicidad de la información, y la influencia del factor humano.

El plan de mejoras trazado debe estar dirigido primeramente a los riesgos identificados para cada dato, de esta forma se evidencia el enfoque a riesgos que se introdujo en el rediseño del procedimiento. Un plan de medidas preventivas para evitar la materialización de riesgos, equivale, al mismo tiempo, a mejoras para la calidad de los datos de la empresa.

➤ Monitoreo.

El monitoreo debe tener, entre sus principales objetivos, el chequeo sistemático de lo establecido en los planes de acción elaborados y aprobados en cada organización. Se debe considerar, como un aspecto, la frecuencia con que se tomarán las observaciones para llevar a cabo el monitoreo en sí.

Para la toma de las observaciones, se tomarán la frecuencia determinada para detectar los errores asociados a cada dimensión y tipo de dato, y calcular los indicadores establecidos; de esta forma, se pudieran establecer gráficos de control para la cantidad de errores o para los indicadores

porcentuales (una vez que la cantidad de datos disponibles lo permita), de forma tal que se pueda observar el comportamiento cronológico de los mismos y determinar si hay o no control estadístico, lo que indicaría el tipo de acciones a realizar.

### **Impacto.**

Para la determinación del impacto del procedimiento, Vilalta (2008) propone la aplicación de una encuesta que pretende medir la percepción del grupo de trabajo sobre determinados aspectos como son: su impacto en la imagen de la empresa, la resolución de problemas asociados a la calidad de los datos o la mejora de las relaciones con los clientes.

En la misma se le solicita a los encuestados que marquen en una escala de 2 a 5 su grado de acuerdo con la afirmación planteada, donde 5 significa total grado de acuerdo, 4 acuerdo, 3 desacuerdo, y 2 desconocimiento sobre el aspecto en cuestión.

Además, como parte del rediseño, se propone medir el costo, que tiene para la empresa, en el hecho de trabajar con datos de mala calidad, haciendo uso de la siguiente ecuación:

$$CT = \sum P_i * N_i * C_i$$

Donde:

CT: costo total.

P<sub>i</sub>: probabilidad de que el dato i tenga errores.

N<sub>i</sub>: cantidad de veces que se emplea el dato i.

C<sub>i</sub>: costo asociado a que el dato i tenga errores.

### **Aplicación del procedimiento en una oficina de administración tributaria.**

Este procedimiento se aplicó en una oficina de administración tributaria, donde los servicios que se brindan van desde la atención, orientación y asesoría personal, por teléfono, correo postal y electrónico, hasta la entrega gratuita de documentos fiscales y programas que calculan los

tributos, documentos de asistencia que indican a cada contribuyente sus obligaciones, y cuándo cumplirlas, además de la publicación de informaciones de trascendencia tributaria por diferentes vías: cartas, sueltos, folletos y medios masivos de comunicación.

Entre sus objetivos se encuentra brindar una atención personalizada a sus clientes, por lo que reconocen la necesidad de trabajar con datos de buena calidad en los procesos de toma de decisiones. Por ello se decide aplicar en esta entidad el procedimiento propuesto, cuyos resultados se exponen a continuación.

### **Etapas preliminar.**

Es necesario el cumplimiento de una serie de requerimientos por parte de la entidad en cuestión. Uno de ellos es la concesión de las personas al equipo de trabajo que serán los responsables de la aplicación del procedimiento, los que se desempeñarán como especialistas. Este grupo fue integrado por un total de 5 expertos.

Por otra parte, la dirección tiene la responsabilidad de asignar los recursos necesarios para la aplicación efectiva del procedimiento. Entre estos recursos se encuentra el tiempo necesario para emplear en esta actividad, tanto del equipo de trabajo que aplicará el procedimiento como de aquellas personas que por su desempeño se verán involucradas. Se deberá permitir el acceso a las bases de datos, teniendo en cuenta que se cumplan los requisitos de seguridad establecidos. Además, la capacitación se encuentra presente a lo largo de todo el procedimiento y es de gran importancia para lograr resultados satisfactorios. Su objetivo es transmitir los conocimientos sobre la necesidad de trabajar con datos de buena calidad y los relacionados con la aplicación del procedimiento.

### **Etapas 1: Planificación**

Para darle cumplimiento a esta etapa se aplican una serie de actividades que se muestran a continuación.



➤ Identificar datos críticos.

La identificación de los datos críticos toma como base un listado confeccionado por el grupo de trabajo de la entidad, el cual contaba con un total de 43 posibles datos a considerar como críticos, ya que tributan a indicadores importantes para la toma de decisiones.

Para cada uno de estos datos se identificaron los posibles riesgos asociados a ellos, teniendo en consideración que en algunas ocasiones un mismo riesgo podía afectar diferentes datos, se determinaron un total de 40 riesgos asociados a los datos objeto de estudio. Posteriormente se analizó el comportamiento de los aspectos propuestos (frecuencia, impacto y eficacia de los controles) en cada caso.

- Frecuencia de los eventos:

El grupo de trabajo de la empresa facilitó la información acerca de la periodicidad de ocurrencia de los riesgos identificados en el marco de trabajo. Esta actividad arrojó como resultado que el 51,35 % de los riesgos analizados se consideran de frecuencia Alta, el 18,91 % de frecuencia Media y el resto de frecuencia Baja.

- Impacto de la materialización de los eventos:

De igual manera, se procedió con la clasificación de las consecuencias o impacto de la materialización de los riesgos en: Catastrófico, Moderado y Leve, resultando:

- Impacto catastrófico: 89,19 % de los riesgos analizados.
- Impacto moderado: 10,81 %.
- Impacto leve: 0 %.

- Detección de los eventos a través de la eficacia de los controles:

La empresa actualmente aplica controles únicamente a aquellos riesgos que con anterioridad se hayan materializado, por tanto no existen controles establecidos para todos los riesgos existentes.

Para estos casos se consideró que la eficacia de los controles era baja. En estas condiciones se puede plantear que de los riesgos analizados el 32,43 % se consideran de alta detección, el 16,22 % de detección media, y el 51,35% de detección baja.

Teniendo en cuenta esos resultados, se calcula el Índice de Riesgo del Dato (IRD), el que arrojó valores entre 5 y 125. Los valores del IRD fueron ordenados de mayor a menor, resultando un total de: 9 riesgos inaceptables, 13 riesgos importantes, 13 riesgos moderados, 5 riesgos tolerables y 0 riesgos aceptables.

Los riesgos inaceptables están asociados a 24 datos, los que se deben considerar como críticos en este estudio. Por interés de la entidad y por cuestiones de tiempo, no es posible analizar todos estos datos, por lo que se seleccionaron aquellos que para la empresa presentaban más problemas en la actualidad a fin de ilustrar la correcta aplicación del procedimiento propuesto.

Los datos finalmente seleccionados en el análisis son: (1) Cantidad de personas naturales inscritas, (2) Cantidad de personas jurídicas inscritas, y (3) Deuda al cierre del mes para personas naturales.

➤ Determinar dimensiones.

Para lograr determinar las dimensiones a utilizar se confeccionó una lista inicial, comenzando por las dimensiones más importantes, siendo estas: actualidad, exactitud e integridad. Posteriormente se decidió, por parte del grupo de trabajo, incorporar otras tales como confiabilidad y consistencia, que a criterio de estos expertos eran relevantes para el diagnóstico. Cada uno de los expertos ponderó las 6 dimensiones que aparecen en la lista, siendo el valor 1 el más importante.

Posteriormente se utilizó el software Minitab para conocer si existe concordancia entre los expertos. En la salida del software se aprecia que el coeficiente de Kendall es igual a 0.975. Al

considerarse este valor cercano a 1 se puede afirmar que hay concordancia entre los expertos. A continuación se muestra la salida del Minitab:

Coefficiente de concordancia de Kendall.

**Coef Chi-cuad. GL P**

0.975 24.375 5 0.0002

De forma complementaria a este procedimiento, se analizó si esta concordancia entre los expertos es significativa o no. Para esto se recurrió a la prueba de hipótesis para la significación del coeficiente de Kendall:

$H_0$ : No hay concordancia  $H_1$ : Hay concordancia RC:  $P\text{valor} \leq \alpha$   $\alpha = 0,05$

Como se muestra en la salida del Minitab, el Pvalor es igual a 0.0002. Esto significa que la región crítica se cumple, se rechaza  $H_0$ , y se puede afirmar que esta concordancia es significativa.

La siguiente tabla muestra las ponderaciones otorgadas por los expertos a cada dimensión en cuestión, el Rj (Rango) y el orden de las dimensiones a utilizar. Se acata a la condición de que el Rj sea menor o igual que el Rj medio para la selección de estas. Para la determinación del orden se tomó como más significativo aquel que tenga el menor Rj.

**Tabla 5:** Ponderación de las dimensiones por los expertos.

<b>Dimensión</b>	<b>E1</b>	<b>E2</b>	<b>E3</b>	<b>E4</b>	<b>E5</b>	<b>Rj</b>	<b>Orden</b>
Actualidad	1	1	2	1	2	7	1
Exactitud	2	2	3	2	2	11	2
Puntualidad	6	5	6	6	6	29	
Integridad	3	3	3	3	4	16	3
Confiabilidad	4	4	5	4	5	22	
Consistencia	4	5	5	4	5	23	

**Fuente:** Elaboración propia.

➤ Determinar indicadores.

Se establecieron un total de 3 indicadores porcentuales para evaluar la calidad de los datos asociados a las dimensiones seleccionadas, los que se muestran a continuación:

- Porcentaje de datos íntegros =  $(\text{datos completos} / \text{total de datos}) * 100$
- Porcentaje de datos exactos =  $(\text{datos exactos} / \text{total de datos}) * 100$
- Porcentaje de datos actualizados =  $(\text{datos actualizados} / \text{total de datos}) * 100$

Teniendo en cuenta las ponderaciones de las dimensiones dadas por los expertos, se propone calcular un coeficiente ponderado, llamado Índice Ponderado de Calidad del Dato (IPCD). El cálculo de este índice permite realizar una valoración del nivel de calidad de los datos de manera más exhaustiva, en la que se incluye la situación de cada una de las dimensiones consideradas.

### **Etapas 2: Ejecución y control.**

Posteriormente se aplica la etapa 2 del procedimiento y sus respectivas actividades.

- Localizar la fuente asociada al dato.

En el caso de los dos primeros datos, la información procede de una aplicación informática con base de datos Oracle y lenguaje de programación Visual Basic. Para el dato 3, la información es recogida en un Excel en dependencia del nivel del que se trate (municipal, provincial o nacional).

Se considera como la fuente del dato, el momento en que se originan cada uno de éstos, en sus respectivos formatos.

- Tomar observaciones.

Por ser de interés para la empresa, se analizará la información de los años 2014 y 2015 en cada una de las provincias, lo que representa un total de 34 datos en cada caso (16 provincias cada año y un resumen anual). No se considera pertinente tener en cuenta información previa a esta fecha, debido a que el organismo no analizaba la información según este formato, motivo por el cual los registros empleados con anterioridad sufrieron modificaciones, y por ende, la información solicitada.

El software SAMPLE permitió determinar, que para una población de tamaño 34, con un nivel de confianza de 95 % y una precisión del 5 %, se debe analizar una muestra de 31 datos. Teniendo en cuenta que no existe mucha diferencia entre el tamaño de muestra necesario y toda la población, se decide trabajar con esta última sin necesidad de aleatorizar los datos.

➤ Calcular indicadores.

Luego de consultar cada una de las fuentes, se describe el error detectado por cada dato y la dimensión que se afecta, tal como se muestra en la tabla 6.

**Tabla 6:** Clasificación de los errores detectados por cada dato.

<b>Dato</b>	<b>Descripción del error</b>	<b>Dimensión afectada</b>	<b>Clasificación</b>
Cantidad de personas naturales inscritas	<ul style="list-style-type: none"> <li>• Desactualizado</li> <li>• No coincide con la fuente</li> </ul>	<ul style="list-style-type: none"> <li>• Actualidad</li> <li>• Exactitud</li> </ul>	Grave Grave
Cantidad de personas jurídicas inscritas	<ul style="list-style-type: none"> <li>• Desactualizado</li> <li>• No coincide con la fuente</li> </ul>	<ul style="list-style-type: none"> <li>• Actualidad</li> <li>• Exactitud</li> </ul>	Grave Grave
Deuda al cierre del mes	<ul style="list-style-type: none"> <li>• Desactualizado</li> <li>• No coincide con la fuente</li> </ul>	<ul style="list-style-type: none"> <li>• Actualidad</li> <li>• Exactitud</li> </ul>	Grave Grave

**Fuente:** Elaboración propia.

Como se puede apreciar, las dimensiones exactitud y actualidad se encuentran afectadas en todos los datos analizados, mientras que no se detectaron errores asociados a la dimensión integridad.

En cuanto a la clasificación de la gravedad de los errores detectados, se consideran todos como graves, debido a que son datos que se emplean por los altos directivos en la toma de decisiones y existen otros datos cuyo procesamiento depende de estos tres que se analizan.

En la tabla 7 se muestra un resumen del comportamiento de cada dimensión, el que permite calcular los indicadores antes mencionados.

**Tabla 7:** Comportamiento de las dimensiones.

Dato	Integridad	Actualidad	Exactitud	IPCD
Cantidad de personas naturales inscritas	34 datos (100 %)	23 datos (67.6 %)	14 datos (41.2 %)	30.6
Cantidad de personas jurídicas inscritas	34 datos (100 %)	32 datos (94.1 %)	29 datos (85.3 %)	45.4
Deuda al cierre del mes	34 datos (100 %)	16 datos (47.1 %)	27 datos (79.4 %)	38.3
Total	102 datos (100 %)	71 datos (69.6 %)	70 datos (68.6 %)	38.1

**Fuente:** Elaboración propia.

$$W_{\text{integridad}} = [5(3+1)-16]/34 = 0.1176$$

$$W_{\text{exactitud}} = [5(3+1)-11]/34 = 0.2647$$

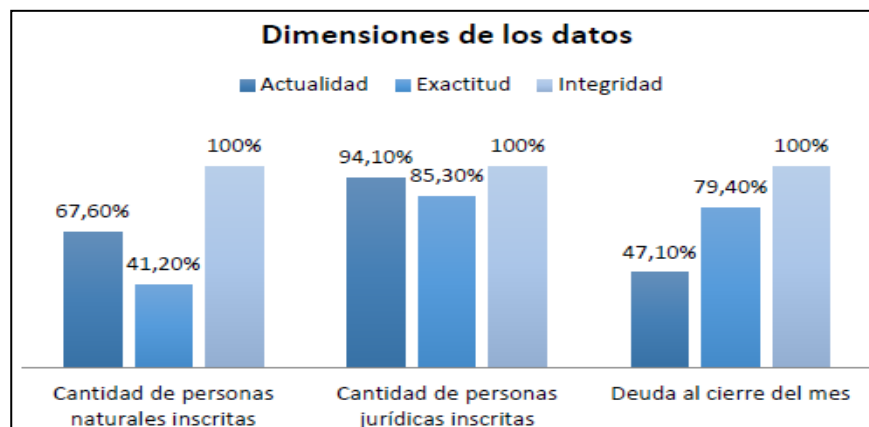
$$W_{\text{actualidad}} = [5(3+1)-16]/34 = 0.1176$$

$$\text{IPCD} = 1(0.1176) + 0.696(0.1176) + 0.686(0.2647) = 0.381$$

➤ Elegir y desarrollar herramientas.

### Dimensiones

Primeramente se construyó un gráfico de barras tal y como se muestra en la siguiente figura, que permite una mejor visualización de los indicadores obtenidos referidos a las dimensiones reflejadas en la tabla 7.

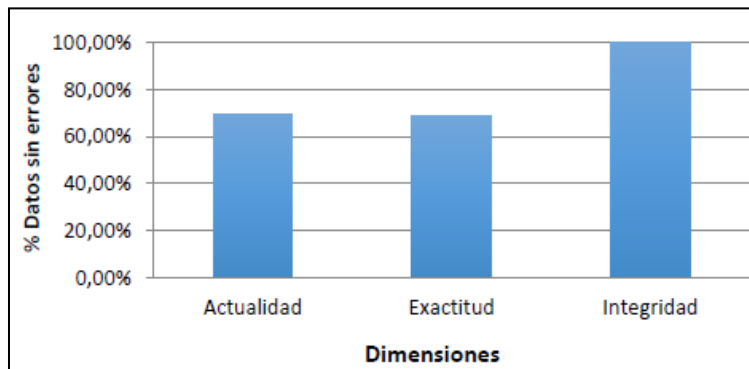


**Figura 2:** Resultados de los indicadores.

**Fuente:** Elaboración propia.

Como se puede apreciar las dimensiones más afectadas son la exactitud y la actualidad en los datos 1 y 3 respectivamente.

Otro análisis que se propone tiene en cuenta los valores promedios de los porcentajes de los errores, tal y como se muestra a continuación:



**Figura 3:** Comportamiento de las dimensiones.

**Fuente:** Elaboración propia.

Comparando los valores promedios de los porcentajes de los errores para cada dimensión estudiada, se puede apreciar claramente que los mayores problemas los presenta la Exactitud y la Actualidad con 68,63% y 69,60 % respectivamente.

#### Datos.

La tabla 8 resume la cantidad de errores que se detectaron por cada dato asociado a las dimensiones.

**Tabla 8:** Cantidad de errores por dato asociados a las dimensiones.

Dato	Dimensión	Errores	%	Sin errores	%
Cantidad de personas naturales inscritas	Actualidad	11	32.35	23	67.65
	Exactitud	20	58.82	14	41.18
Cantidad de personas jurídicas inscritas	Actualidad	2	5.88	32	94.12
	Exactitud	5	14.71	29	85.29
Deuda al cierre del mes	Actualidad	18	52.94	16	47.06
	Exactitud	7	20.59	27	79.41

**Fuente:** Elaboración propia.

Se puede apreciar que los principales problemas los presenta la dimensión Exactitud para el dato 1 (58,82 % de los datos están afectados en esta dimensión), lo mismo ocurre para el dato 2 (14,71 % de los datos afectados por esta dimensión), y para el dato 3, la dimensión más afectada es la actualidad (52,94 % de los datos afectados por esta dimensión).

### **Etapas 3: Mejoramiento.**

En este apartado se aplica la etapa 3 del procedimiento con cada una de las actividades que lo conforman.

➤ **Análisis de los problemas y mejora.**

Para determinar las causas fundamentales que ocasionan estos errores, se realizó una tormenta de ideas por parte de los expertos que han intervenido a lo largo del procedimiento. Se identificaron un total de 5 causas, las que se relacionan a continuación:

- No se realiza una debida revisión sistemática del registro de contribuyentes, lo que provoca que estos sean recibidos con errores en los datos.
- No existe un mecanismo para garantizar que se envíe la información en tiempo, las medidas que se toman (enviar avisos o requerimientos vía e-mail) no son efectivas.
- Falta de preparación, motivación e inadecuadas condiciones de trabajo del personal encargado de la entrada de datos.
- No se cuenta con una base de datos nacional digital y automatizada que permita detectar la duplicidad de datos.
- Debido a la alta fluctuación de la fuerza laboral, la baja calificación y la cuantiosa carga de trabajo, ocurren muchos errores en la captación de datos, que en ocasiones no son detectados y por tanto no se corrigen.



➤ Monitoreo.

Después de ser detectados y analizados los problemas y sus causas, se procedió a la elaboración de un plan de acciones encaminado a la erradicación de estos problemas de calidad de los datos, partiendo desde evitar la materialización u ocurrencia de sus riesgos.

*Acciones que conforman el plan:*

1. Realizar encuentros con los directivos con el objetivo de que entiendan la importancia que tiene este proceso.
2. Amonestar de forma más severa los incumplimientos con los términos de entrega de la información requerida.
3. Mantener una revisión sistemática al registro de contribuyentes, en especial, a los documentos que presentan los mismos en el momento de la inscripción.
4. Lograr establecer un único Sistema de Información Automatizado que elimine la duplicidad de los datos y que optimice el trabajo.

Para cada una de estas acciones se define fecha de ejecución, responsable y participantes.

**Impacto.**

Para medir el impacto que tuvo el procedimiento se aplica una encuesta a las personas involucradas en el estudio, cuyos resultados se muestran a continuación. El análisis se realiza a partir de la moda en cada caso, que es una medida bastante utilizada en estudios de este tipo (Vilalta, 2008).

**Tabla 9:** Resultados de la encuesta.

Aspecto a responder	Puntuación			
	5	4	3	2
1. La aplicación del procedimiento es efectiva para la detección de errores y problemas de calidad de los datos.	100			
2. La aplicación del procedimiento facilita la toma de acciones dirigidas a resolver los problemas con la calidad de los datos.	100			
3. La aplicación sistemática del procedimiento permitirá resolver a mediano plazo (2 años) los problemas de calidad de los datos.			100	
4. La aplicación del procedimiento permite reducir las pérdidas económicas en las que se incurre por concepto de mala calidad de los datos.		60		
5. La aplicación del procedimiento mejora la satisfacción de los trabajadores usuarios de los datos dentro de la organización.			60	
6. La aplicación del procedimiento permite mejorar las relaciones con los clientes.			100	
7. La mejora de las relaciones con los clientes facilitará la captación de nuevos clientes.			100	
8. La aplicación del procedimiento beneficia el establecimiento y puesta en práctica de la estrategia organizacional.			100	
9. La aplicación del procedimiento permite conocer otros problemas organizativos no relacionados con la calidad de los datos.				60

**Fuente:** Elaboración propia.

Como se puede apreciar, la aplicación de la encuesta permite corroborar que la aplicación del procedimiento tiene una gran influencia en los procesos que se vinculan directamente con la calidad de los datos, destacándose las facilidades que brinda para la detección de errores y problemas de calidad de los datos, la toma de acciones dirigidas a resolver estos problemas, y las pérdidas económicas en las que se incurre por este concepto.

Por otra parte, para calcular el costo total es necesario determinar el costo asociado a que cada dato tenga errores ( $C_i$ ). En los casos analizados este valor se puede determinar con las siguientes expresiones:

- Dato 1: Cantidad de personas naturales (PN) inscritas.

$C1 = \text{Promedio de PN sin inscribir} \times \text{Recaudación anual promedio por PN inscrita.}$

$$C1 = 6 \text{ personas} \times 1\,952,1 \text{ MP/personas-año.}$$

$$C1 = 11\,712,6 \text{ MP/año.}$$

- Dato 2: Cantidad de personas jurídicas (PJ) inscritas.

$C2 = \text{Promedio de PJ sin inscribir} \times \text{Recaudación anual promedio por PJ inscrita.}$

$$C2 = 47\,909 \text{ personas} \times 2,6 \text{ MP/personas-año.}$$

$$C2 = 125\,842,2 \text{ MP/año.}$$

Para el dato 3 (deuda al cierre del mes) se asume, como este valor, la cantidad promedio de dinero que se dejó de reportar (diferencia entre la fuente del dato y aquel que se emplea en la toma de decisiones). En este caso, este valor es de 79 089 MP.

La probabilidad de que los datos tengan errores se asume como el porcentaje de veces que el dato no es exacto, ya que esta es la dimensión que expresa los errores referidos a la precisión de los datos (según los errores analizados en este caso).

Luego, el costo total, según la expresión presentada en la descripción del procedimiento, sería:

$$CT = \sum P_i * N_i * C_i$$

$$CT = (0,59 \times 1 \times 11\,712,6) + (0,85 \times 1 \times 125\,842,2) + (0,21 \times 1 \times 79\,089)$$

$$CT = 130\,484,994 \text{ MP}$$

Esto indica el costo total que representa para la entidad objeto de estudio el trabajar con los datos analizados cuando éstos presentan mala calidad, valor que asciende aproximadamente a 130 484,994 MP al año.

## **CONCLUSIONES.**

Como conclusiones del trabajo de investigación se plantea que:

- La primera etapa del procedimiento se rediseñó siguiendo el enfoque basado en riesgos de las Normas ISO 9001:2015, NC ISO 31000:2015 y NC ISO 31010:2015.
- Se diseñó una herramienta que permite cuantificar lo que implica para una empresa el trabajar con datos con mala calidad, teniendo en cuenta el costo asociado a que el dato tenga error y la probabilidad de ocurrencia del mismo.
- La aplicación del procedimiento ofrece determinadas ventajas para la entidad, destacándose las facilidades que brinda para la detección de errores y problemas de calidad de los datos, y la toma de acciones dirigidas a resolver estos problemas y las pérdidas económicas en las que se incurre por este concepto.
- Se determinó el costo total que representa para la entidad objeto de estudio el trabajar con los datos analizados cuando éstos presentan mala calidad, valor que asciende aproximadamente a 130 484,994 MP al año.

## **REFERENCIAS BIBLIOGRÁFICAS.**

1. Bianchi, B. V. M. (2009). Un caso de estudio en calidad de datos para ingeniería de software empírica. Universidad de la República, Uruguay.
2. Calero, C. (2007). Normalización de la IS: el papel de la universidad. Trabajo presentado en la Universidad de Castilla-La Mancha, España.
3. Haro, F. (2007). Estudio sobre calidad de datos. Consultado 24-02-2016, disponible en <http://www.felixharo.es/estudio-sobre-calidad-de-datos/>
4. Norma ISO 9001:2015. Sistemas de gestión de la calidad — Requisitos.

5. Ortuño, M. (2005). Besterfield. Control de calidad. Consultado 20-11-2015, disponible en <http://es.slideshare.net/margelisortuno/besterfield-control-de-calidad>
6. Redman, T. (2004). Data: An unfolding quality disaster. Consultado 23-01-2016, disponible en <http://www.dmreview.com/portals/dataquality>.
7. Vilalta, J.A. (2008). Procedimiento para el diagnóstico de la calidad de los datos en organizaciones cubanas. Tesis presentada en opción al grado científico de Doctor en Ciencias Técnicas. Instituto Superior Politécnico "José Antonio Echeverría", CUJAE. La Habana, Cuba.

#### **DATOS DE LOS AUTORES.**

1. **Liset Agüero Zardón.** Ingeniera Industrial y profesora de la Facultad de Ingeniería Industrial del Instituto Superior Politécnico "José Antonio Echeverría". Correo electrónico: [lagueroz@ind.cujae.edu.cu](mailto:lagueroz@ind.cujae.edu.cu)
2. **José Alberto Vilalta Alonso.** Ingeniero Industrial y Doctor en Ciencias Técnicas. Profesor Titular de la Facultad de Ingeniería Industrial del Instituto Superior Politécnico "José Antonio Echeverría". Correo electrónico: [jvilalta@ind.cujae.edu.cu](mailto:jvilalta@ind.cujae.edu.cu)
3. **Rosario Caridad Garza Ríos.** Ingeniera Industrial y Doctora en Ciencias Técnicas. Profesora Titular de la Facultad de Ingeniería Industrial del Instituto Superior Politécnico "José Antonio Echeverría". Correo electrónico: [rosariog@ind.cujae.edu.cu](mailto:rosariog@ind.cujae.edu.cu)
4. **Idalianys Urquiola García.** Ingeniera Industrial y profesora de la Facultad de Ingeniería Industrial del Instituto Superior Politécnico "José Antonio Echeverría". Correo electrónico: [iurquiolag@ind.cujae.edu.cu](mailto:iurquiolag@ind.cujae.edu.cu)

**RECIBIDO:** 5 de septiembre del 2016.

**APROBADO:** 29 de septiembre del 2016.