Orientando la toma de decisiones con datos en el sector público

Un enfoque práctico

David Chaves Campos, Elzbieta Malinowski
Escuela de Ciencias de la Computación e Informática
Universidad de Costa Rica
San Pedro de Montes de Oca, Costa Rica
{david.chavescampos, elzbieta.malinowski}@ucr.ac.cr

***Resumen*—Tener datos y no aprovecharlos. En una nueva era donde la información es de mucha importancia para tomar mejores decisiones, esto es una situación lamentable. Aun así, es una realidad que afecta muchos ámbitos, como sectores públicos, donde de forma sistemática se recolectan datos valiosos y solo se utilizan para hacer algunas mediciones de interés político. Este trabajo ejemplifica el uso del modelo multidimensional al ser aplicado a un conjunto de datos públicos con el objetivo de mejorar su acceso y ofrecer diferentes posibilidades de análisis para la toma de decisiones. Se presenta el proceso de construcción del sistema y los resultados de diferentes consultas analíticas, permitiendo que cualquier sector pueda utilizarlo con datos de su interés y con ello, mejore la forma en cómo decide. Esto con el objetivo de proporcionar herramientas efectivas que procuren un progreso continuo en la manera mediante la que diversos agentes de la sociedad hacen su trabajo.**

Palabras clave—enfoque multidimensional; toma de decisiones; gobierno; datos públicos; análisis dinámico; encuesta

I. Introducción

Ante el cuestionamiento de si pueden los datos mejorar la calidad de vida de las personas suele tenerse una respuesta afirmativa. Esto porque se vive en una época donde se considera que, si los datos se utilizan de una forma adecuada, pueden orientar de manera efectiva el mejoramiento de procesos que incidan positivamente en el bienestar común. Por ello, en múltiples ocasiones es importante plantearse esta pregunta en escenarios que tienen gran influencia en la vida de las personas, como puede ser el sector público de un país. En este caso, si bien se esperarían resultados igualmente efectivos, lamentablemente no es un enfoque que suela utilizarse.

Por esta razón, se considera de importancia estudiar la forma mediante la cual podría incentivarse una mayor utilización de datos para el soporte en la toma de decisiones en el sector público de Costa Rica, el cual agrupa diversas instituciones del país. Esto debido a que aquellas disposiciones que se tomen con un criterio más informado posibilitan mejorar su alcance y efectividad, lo que supone un beneficio para la colectividad. Para esto, se propone explorar la forma en cómo pueden utilizarse datos de origen público, como los proporcionados por la Encuesta Nacional de Hogares [1] y plantear un modo que incentive su uso en diversos sectores, de una manera simple para usuarios que emplean de forma habitual programas ofimáticos, sin necesidad de sistemas de análisis sofisticados. Por lo anterior, se valora que la utilización de un modelo multidimensional posibilita una apropiada presentación de los datos en aplicaciones de uso habitual, como hojas electrónicas, reduciendo la complejidad en la ejecución de las consultas por parte de los usuarios finales. Esto debido a que es un enfoque que permite visualizar los datos numéricos desde diversas perspectivas mediante el uso de tablas dinámicas. Estas posibilitan construir, por ejemplo, gráficos interactivos que faciliten la interpretación de los datos y puedan conducir a un mejor planteamiento de futuras decisiones, sin necesidad de contar con un elevado conocimiento técnico.

Adicionalmente, al incentivar el uso de este tipo de aplicaciones se promueve que los usuarios tengan una mayor disposición por utilizar los datos de manera cotidiana en el respaldo de sus decisiones. Esto supondría un avance respecto a los posibles análisis que pueden realizarse y la adopción de una posición donde se enfatice la relevancia que tienen los datos para contribuir en el alcance de objetivos importantes para cada sector. Lo anterior se puede considerar, de acuerdo con lo propuesto por la consultora Gartner [2], como un incentivo para alcanzar mayores niveles de madurez en cuando a la adopción de datos para la toma de decisiones. Con esto se hace referencia a que el uso creciente de datos permite mejorar gradualmente la forma en cómo estos pueden respaldar diferentes resoluciones, siendo necesario, por ejemplo, que para utilizar herramientas de análisis dinámico debe existir un conocimiento previo sobre el uso de reportes estáticos en hojas electrónicas.

Asimismo, la propuesta de nuevos enfoques que faciliten el análisis de los datos permite definir posibles aplicaciones como, por ejemplo, la medición del alcance de los programas sociales o el diagnóstico de las condiciones laborales en los diferentes segmentos de la población. Además, estos son instrumentos que contribuyen al alcance de objetivos comunes como las Metas de Desarrollo Sostenible propuestas por las Naciones Unidas. [3] Lo anterior debido a que se considera la utilización de *Big Data* como un factor que contribuye positivamente en su logro, al permitir la monitorización continua de variables asociadas con el bienestar de los habitantes. Esto hace que exista un incentivo por un mayor uso de datos en las diversas actividades de sectores públicos y privados, lo que permita, de acuerdo con los niveles

de madurez propuestos por Gartner [2], adoptar este tipo de tecnologías en el futuro.

Este artículo se encuentra estructurado en siete secciones. La sección II se refiere a la situación actual en el uso de datos en Costa Rica, mientras que III sección presenta el origen de los datos utilizados. El esquema multidimensional, los proceso ETL y la implementación del cubo OLAP se presentan en las secciones IV, V y VI, respectivamente. La sección VII incluye algunos casos de aplicación y finalmente, las conclusiones generales del trabajo se encuentran en la sección VIII.

II. Situación Actual

La Encuesta Nacional de Hogares (ENAHO) corresponde con un programa desarrollado por el Instituto Nacional de Estadísticas y Censos (INEC) [2] desde el año 1976 con el objetivo de determinar el nivel de bienestar de la población de acuerdo con la caracterización de los hogares, el ingreso percibido y su acceso a servicios de seguridad social y educación. Este estudio es aplicado de forma anual durante el mes de julio en todo el territorio de Costa Rica, con una muestra representativa de cerca del 1% del total de la población. Los resultados de esta encuesta son empleados para efectuar investigaciones y diagnosticar la distribución del ingreso en la población mediante la definición de quintiles y el cálculo del coeficiente de Gini; lo que permite establecer el nivel de desigualdad en el nivel de riqueza que existe en la población.

Adicionalmente, se realiza una caracterización de los niveles de pobreza de acuerdo con su ubicación geográfica y en algunas ocasiones se incorporan módulos adicionales relacionados con algún tema específico como la adopción de telecomunicaciones o la utilización de servicios sociales. Estos estudios permiten realizar un seguimiento periódico de algunas variables de interés gubernamental como la pobreza, distribución de la riqueza y desempleo en proyectos como el Plan de Desarrollo elaborado por el Ministerio de Planificación [4]. En general, el análisis efectuado es con un enfoque estadístico que utiliza los datos de una manera agrupada de acuerdo con características existentes previamente como divisiones geográficas o nivel de urbanización. Sin embargo, al agregarse los datos se desaprovecha la granularidad asociada con cada persona y hogar encuestado, lo que podrían brindar otras perspectivas de análisis.

Los resultados de las investigaciones se publican en el sitio de INEC en forma directa mediante hojas electrónicas o documentos *pdf*, o bien, se ofrece sin procesar a personas interesadas para su utilización con software estadístico propietario (*SPSS*). Lo anterior limita el uso que se le pueda dar a la encuesta, ya que, si se utilizan los documentos publicados, se necesitan complejos procesos de integración de datos de acuerdo al interés del investigador. Por otro lado, con los datos suministrados directamente, se requiere un procesamiento previo de estos con el fin de obtener información relevante, lo que en múltiples ocasiones se encuentra fuera del alcance de diversas áreas académicas y sectores sociales, ya que es necesario contar con software estadístico especializado y cierto conocimiento técnico para poder utilizarlo.

Las dificultades en integración o manipulación de los datos motivan a otras instituciones públicas a contar con proyectos que hacen uso de estos datos para ofrecerlos públicamente de una manera diferente. Por ejemplo, el Observatorio del Desarrollo [5], un ente de investigación adscrito a la Universidad de Costa Rica, mediante la iniciativa "Tendencias del Desarrollo Costarricense" [6] y el proyecto "Costa Rica en Cifras" [7] posibilita la visualización de variables como el promedio de hogares de acuerdo a la zona de urbanización y el porcentaje de pobreza mediante gráficos de construcción sencilla. Esto se realiza en su sitio web utilizando la herramienta *Tableau* en un modo de solo lectura, lo que, si bien permite vislumbrar la evolución de las variables seleccionadas en el proyecto, no permite ninguna interacción con el usuario. Como consecuencia, la situación se asemeja con lo presentado por el INEC, aunque en algunas ocasiones se utilizan datos de otras fuentes para plantear otros enfoques.

De forma similar, el Centro Centroamericano de Población (CCP) [8], institución que también se circunscribe dentro de la Universidad de Costa Rica, ofrece el acceso a datos importantes para investigaciones demográficas, como es el caso de la Encuesta Nacional de Hogares [9]. Estos datos se obtienen, mediante un filtrado de las variables requeridas y se permite hacer algunos cálculos sencillos como la ponderación. Para ello, el sitio web utiliza la herramienta PDQ-Explore [10], misma que se encuentra diseñada para obtener tabulaciones de datos demográficos. Esto, si bien permite la extracción de la información, no ofrece características que faciliten el análisis por parte de los usuarios, ya que los datos no se integran entre sí, limitándose solo a la visualización de variables específicas.

III. Origen de los datos para el caso de estudio

Los datos empleados en el presente trabajo corresponden con encuestas realizadas por el INEC entre los años 2010 y 2016. Estos son de origen público y corresponden con datos abiertos que pueden ser utilizados de forma libre siempre que se cumpla con sus términos y condiciones. Los mismos agrupan alrededor de 500 variables por cada registro, aunque dependiendo del año, esta cifra es cambiante en función de las necesidades de investigación. Estos datos se proporcionan en formato *sav*, para su uso en software estadístico, de manera que puede representar una limitante para algunos de los usuarios que los accedan. Además, por la forma en como está construida la encuesta, se presentan dos problemas. El primero corresponde con que, al ser recolectados para elaborar una representación de la población costarricense, los elementos de la muestra son actualizados en un 25% cada año [11]. Eso hace que solo pueda existir seguimiento de un hogar después del período de cuatro años, ya que la muestra es actualizada en su totalidad al finalizar este período de tiempo. El segundo problema corresponde a que, por posibles razones de protección de privacidad de los datos, estos se encuentran identificados por una llave primaria compuesta que no registra relación longitudinal entre los períodos correspondientes. De manera que, por ejemplo, el hogar diez en el período uno tiene una llave desconocida en el período dos. Como consecuencia, los posibles análisis que se puedan realizar para cada hogar en particular se encuentran limitados para cada período en específico. A pesar de esto, si es posible realizar comparaciones al agregar los datos con respecto a algunas características, por ejemplo, por género y región para calcular la evolución de las bonificaciones salariales recibidas por mujeres residentes en la región Central a través de los períodos estudiados.

Estado físico de la vivienda	Variables originales
Malo	Paredes = 1 y Paredes + Techo + Piso = 3, 4, 5
Regular	Paredes = 2 y Techo = 1 ó 2 y Piso = 1 ó 2 ó Paredes = 3 y Techo=1 y Piso=1 ó Paredes + Techo + Piso = 6 ó 7
Bueno	Paredes + Techo + Piso = 8 ó 9

Fig. 1. Construcción de la variable sobre el estado físico de la vivienda

Adicionalmente, dentro de las variables seleccionadas para el análisis presente en este trabajo se procuró utilizar en su mayoría las que son construidas por el INEC. Esto debido a que las mismas agrupan varias respuestas del cuestionario para determinar alguna característica de interés. Por ejemplo, en la figura 1 se muestra la forma como se construye la variable del estado físico de la vivienda de acuerdo con los resultados de las preguntas de la encuesta que describen el estado de las paredes, el techo y el piso de la vivienda. Esta es una característica de utilidad debido a que posibilita clasificar los hogares y establece una condición sobre la calidad de vida bajo la que residen los habitantes del país.

Además, algunas otras variables construidas por la institución corresponden con la calificación del hogar y el quintil. La primera agrupa la evaluación del estado físico, los servicios públicos disponibles y la cantidad de personas que residen en la vivienda. Asimismo, el quintil se calcula con los ingresos de cada persona residente y permite establecer el nivel socioeconómico de cada hogar, con respecto a los demás. Estas características se utilizan considerando que los métodos del INEC para la construcción de las mismas son lo suficientemente robustos de acuerdo con la realidad demográfica del país. Asimismo, las mismas están disponibles para disminuir el volumen de los datos y facilitar su manejo por parte de los usuarios de la encuesta, para quienes podría resultar complejo definir estas variables si se carece de criterio técnico.

IV. Uso del enfoque multidimensional

La creciente necesidad de emplear datos para soportar la toma de decisiones en los sectores público y privado motiva la implementación de un sistema que facilite el acceso a los datos disponibles públicamente. Esto, de manera que se pueda manipular de forma dinámica por medio de herramientas comúnmente utilizadas como lo son las hojas electrónicas, y en el caso particular de este trabajo, *Microsoft Excel*. Con ello, se permitiría que los posibles usuarios puedan acceder a información valiosa sin necesidad de tener un conocimiento avanzado en estadística y en la utilización de herramientas asociadas con esta área. Para esto se emplea el modelo multidimensional.

A. Esquema conceptual

El uso del modelo multidimensional es bien aceptado en las aplicaciones tradicionales referentes a ventas y/o compras de productos o servicios, comportamiento de clientes, turismo, seguros, entre otros, donde los datos se obtienen de los sistemas operacionales o transaccionales [12]. El caso de estudio presente en este trabajo exhibe otro tipo de aplicación, considerando diferentes personas y hogares como el enfoque de análisis y alimentándose con datos tipo *snapshot*, es decir, "foto instantánea" de datos para un momento definido en el tiempo.

La figura 2 presenta un esquema conceptual multidimensional [13] con dos enfoques de análisis representados como rombos y llamados relaciones factuales: "Hogar" y "Persona". De esta forma, se aprovecha el modo como se construye la encuesta, donde en primer lugar se realiza una entrevista sobre los aspectos más importantes del hogar y posteriormente se efectúan cuestionarios individuales a cada miembro perteneciente al mismo. Esto permite que ambos entes puedan ser caracterizados desde dos perspectivas diferentes en función de las particularidades que los delimitan.

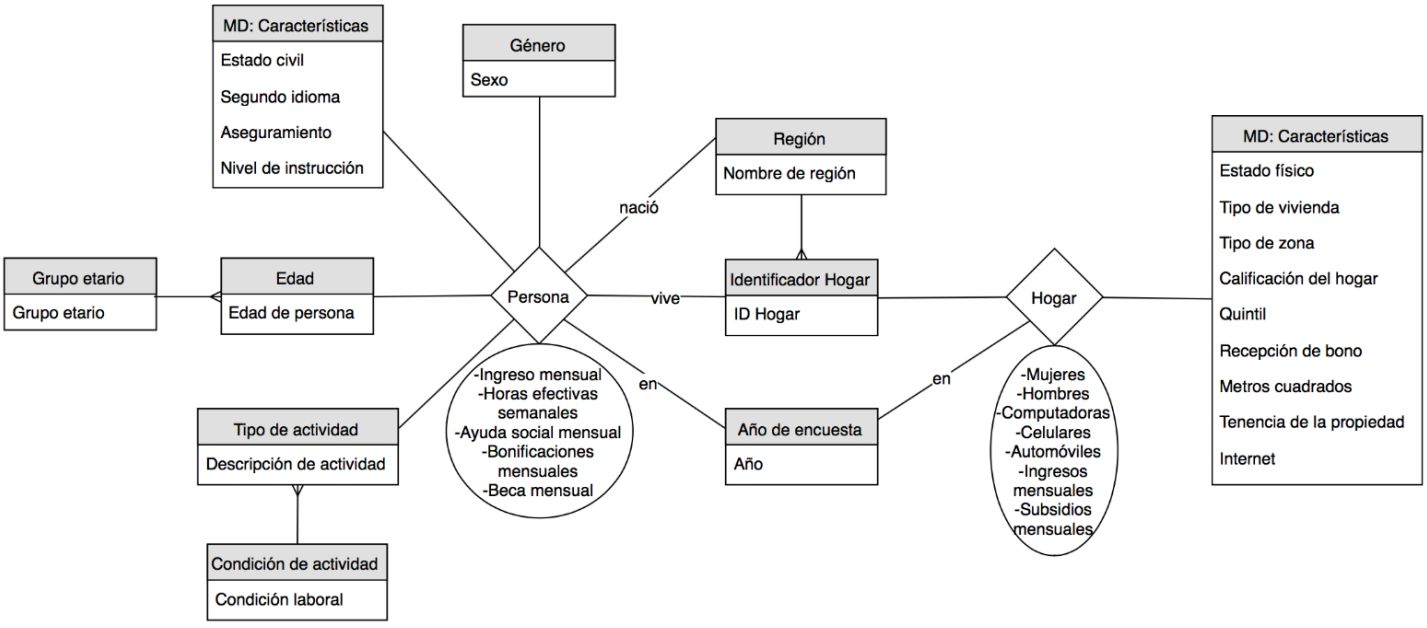

Fig. 2. Esquema conceptual para el análisis de datos demográficos

En el esquema presentado, ambos enfoques de análisis poseen medidas, las cuales se encuentran agrupadas por los óvalos asociados a cada rombo. Estas corresponden con atributos usualmente numéricos que permiten caracterizar a la relación factual de acuerdo con un determinado conjunto de perspectivas. Por ejemplo, la medida "Ingreso mensual" para la relación factual "Hogar" representa el monto total de ingresos para un hogar especifico de acuerdo con sus características físicas.

Como se puede observar en la figura 2, los atributos elegidos como medidas para la relación factual "Persona" posibilitan caracterizar la situación económica de las personas al cuantificarse, de forma mensual, su ingreso principal, posibles bonificaciones, así como ayudas sociales y becas recibidas mediante transferencias de dinero. Además, se incluye la cantidad de horas que trabajó cada individuo de forma semanal, como una métrica sobre su condición laboral.

En el caso del enfoque de análisis correspondiente al "Hogar" se emplean medidas que pueden clasificarse en tres categorías diferentes: la relacionada con la cantidad hombres y mujeres residentes en el hogar, la relativa con los ingresos y subsidios percibidos de forma mensual y la que agrupa algunos bienes con los que cuenta la vivienda, tales como los automóviles y artículos electrónicos; siendo la última medida una que permite conocer si sus habitantes adoptan un perfil tecnológico singular. Asimismo, para ambos enfoques de análisis se utilizan medidas aditivas, de manera que, al agruparse los datos estas pueden ser sumadas entre sí, lo que permite, por ejemplo, calcular el total de mujeres que hay en una zona específica al agregarse esta medida para todos los hogares de esa región. Además, es necesario tener en cuenta que la granularidad de las medidas, es decir, su nivel de detalle, no es uniforme con respecto a la unidad de tiempo de la encuesta, realizada de forma anual. Esto no representa un problema al analizar los datos dentro de un mismo período porque constituyen información para un momento específico en el tiempo. Aun así, en los escenarios que involucran diferentes períodos, es necesario anualizar estas medidas, ya que, de lo contrario, estas no podrían ser aditivas al tener una unidad de tiempo diferente.

De forma similar, en la figura 2 se puede observar que cada enfoque de análisis tiene asociadas dimensiones representadas en forma de rectángulos. Las dimensiones constituyen las perspectivas desde las cuales es posible analizar a determinada relación factual, por ejemplo, usando la dimensión "Género" se puede distinguir entre las horas trabajadas por hombres y mujeres. Para la relación factual "Persona" se utilizan dos tipos de dimensiones: uno que agrupa las características que no cambian con el tiempo y otro que incorpora las que sí pueden presentar esta variación. En el primer escenario se agrupan dimensiones con un único atributo correspondientes con el "Género", la "Región de nacimiento" y el "Año de encuesta". El hecho de que solo incluya un atributo se realiza a nivel conceptual con el objetivo de facilitar al usuario la forma en cómo se pueden agrupar los datos. Asimismo, el "Género" y la "Región de nacimiento" corresponden con características que son intrínsecamente definidas para cada persona, además de que, en el caso del género, la regulación costarricense no considera un posible cambio. En el caso de la dimensión de "Año de encuesta", esta se utiliza para realizar el cambio entre cada una de las encuestas o para dar un seguimiento interanual de los datos agrupados por alguna característica, por ejemplo, al asociarlos por "Región" o "Género".

Un segundo tipo de dimensión empleada corresponde con la, así llamada mini-dimensión [12], misma que facilita la implementación del modelo multidimensional sin incurrir a las comúnmente denominadas *slowly-changing dimensions* (dimensiones que cambian lentamente), proporcionando una forma apropiada de agrupar un acervo de características cambiantes en el tiempo. Estas se encuentran representadas en la figura 2 con los mismos rectángulos de las dimensiones, incluyendo un MD al inicio de su nombre. Las mismas son construidas al definir todas las posibles combinaciones existentes entre diferentes valores de los atributos presentes en la mini-dimensión, es decir, creando el producto cartesiano entre dominios de los atributos que forman esta dimensión. Para caracterizar a cada persona esta se asocia a un registro de esta dimensión, siendo posible que una entrada aplique para más de una persona. Por ejemplo, un registro particular correspondiente con una persona casada, con dominio del idioma francés, asegurada mediante sus cotizaciones laborales y que ha alcanzado un nivel de instrucción de secundaria puede asociarse con un hombre o una mujer de edades y zonas distintas. De esta forma, si se requiere seguimiento de una persona en particular, podría ocurrir que en el siguiente período cambie su estado conyugal a viudo; para representar este cambio se le asignaría una nueva entrada que considere esa nueva combinación.

Adicionalmente, dentro de las dimensiones que cambian a través del tiempo, se consideran diferentes casos donde se establece una jerarquía que permite agrupar los datos de acuerdo con alguna característica en particular. Esto es representado mediante dos rectángulos llamados niveles relacionados entre sí, como se puede ver en la figura 2 en el caso de "Edad" y "Grupo etario". Con las jerarquías es posible sumar las medidas de las personas pertenecientes al mismo grupo de edades, por ejemplo, entre 10 y 14 años. Este grupo de edad representa un nivel no inherente a los datos que se crea con el propósito de clasificar las personas con una edad específica dentro de diferentes grupos etarios. Por ejemplo, en la infancia se agrupan las personas con edades menores o iguales a cuatro años, mientras que la adultez mayor comprende al grupo de personas con al menos 60 años de edad. La definición de esta jerarquía permite realizar un análisis agregado con respecto a las condiciones laborales y económicas dentro de las que es posible ubicar a cada grupo en particular. Por ejemplo, se esperaría que los grupos etarios de adolescencia y juventud se encuentren estudiando y sin reportar rentas, además de que en las etapas la adultez se presume una concentración de la fuerza laboral del país y, por lo tanto, la que genere la mayor parte de los ingresos.

La otra jerarquía asociada con la relación factual "Persona" se refiere al tipo de actividad que desempeña la persona. La misma corresponde con una jerarquía natural, es decir, que es inherente al atributo, por lo que al definirse lo que determinado individuo realiza es posible establecer la condición de actividad que tiene. Por ejemplo, si una persona es estudiante de primaria o secundaria ("Tipo de actividad"), es posible agruparlo como estudiante ("Condición de actividad"), que es una clasificación relevante para efectos de explicar, por ejemplo, una posible nulidad de ingresos.

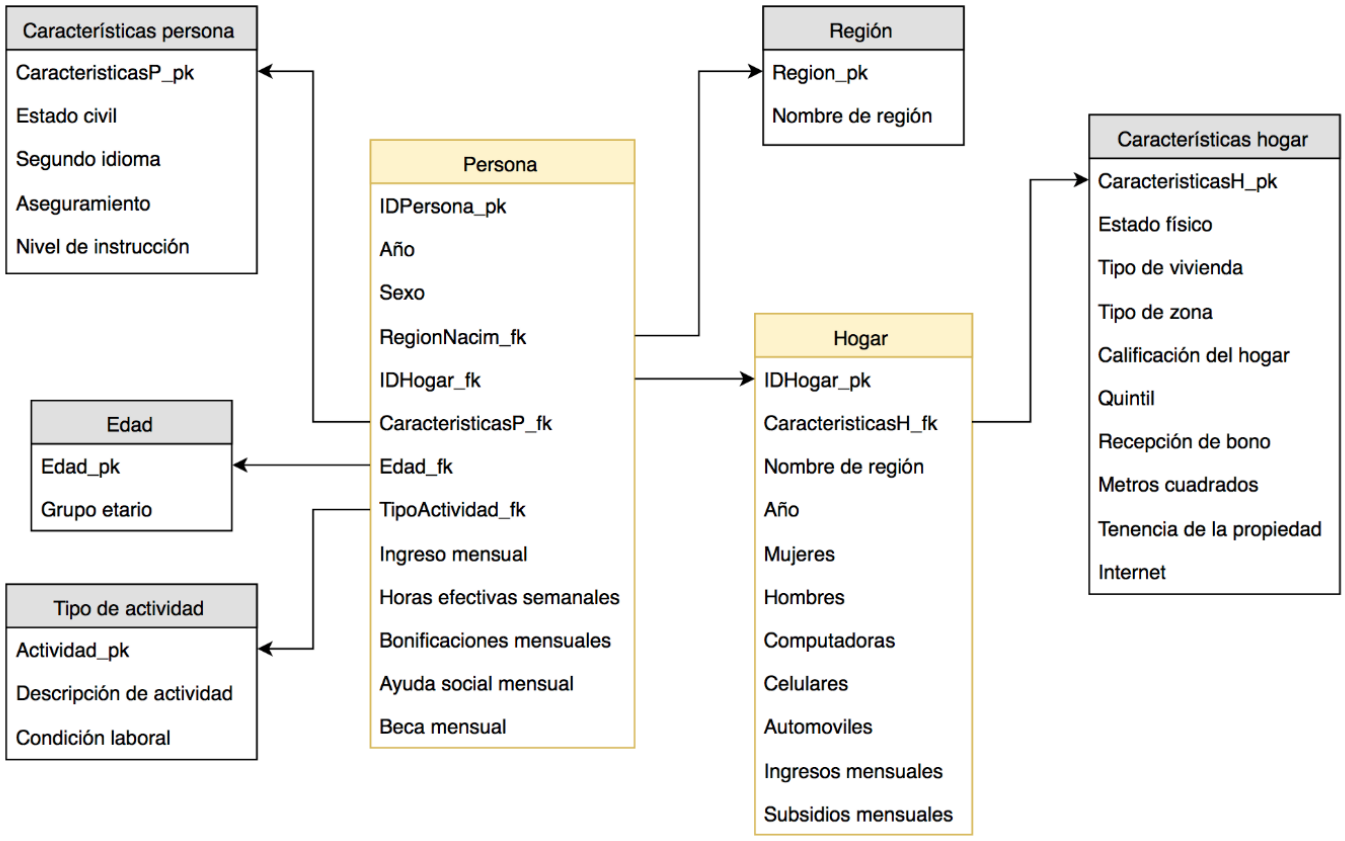

Fig. 3. Esquema lógico de implementación del modelo multidimensional

Con respecto al segundo enfoque de análisis, “Hogar” en la figura 2, es posible notar que se comparten las dimensiones “Año de encuesta” e “Identificador Hogar”. La primera cumple una función análoga al enfoque “Persona”, mientras que la dimensión “Identificador Hogar” se considera del tipo degenerada [12]. Esta clasificación permite agrupar los miembros de un mismo hogar, de forma similar a como en un modelo de ventas, se utiliza el número de factura para asociar los ítems correspondientes a una misma venta. Además, para esta dimensión, se utiliza una jerarquía natural de ubicación geográfica que posibilita la agrupación de los datos de acuerdo con las regiones donde se encuentra ubicada la vivienda. Similarmente, se puede observar el uso de una mini-dimensión para el enfoque relacionado con el hogar, mismo que permite representar con más facilidad los cambios que pueden ocurrir en los hogares a través del tiempo, de una forma equivalente a lo planteado para este tipo de dimensión en el enfoque “Persona”.

B. Esquema de implementación

Para la implementación de este modelo se siguen las reglas de mapeo para un modelo multidimensional propuestas por Malinowski y Zimányi [13], quienes plantean que cada nivel del este modelo corresponde a un tipo de entidad en el modelo entidad relación (ER), el tipo de relación entre niveles que forman jerarquía corresponde con un tipo de relación binaria con cardinalidad 1:n entre estos niveles y la relación factual corresponde a un tipo de relación n-aria del modelo ER. De esta manera se pueden aplicar las reglas tradicionales de mapeo del modelo ER al modelo relacional y obtener como resultado el esquema de implementación (o lógico) expuesto en la figura 3.

Para este escenario es necesario tener en cuenta que para el caso de las dimensiones “Sexo” y “Año de encuesta”, estas son incluidas dentro de la tabla correspondiente a la relación factual, llamada tabla de hechos y no como tablas individuales. Esto se hace debido a que resulta ineficaz construir tablas que contienen un único atributo. Un caso similar ocurre con las dimensiones que se asocian con el identificador del hogar y su ubicación geográfica, mismos que son incorporadas de manera factual dentro de la tabla de hechos “Hogar”. Adicionalmente, se debe considerar que para las jerarquías relacionadas con las dimensiones “Edad”, “Tipo de actividad” y “Región” los datos se incorporan dentro de una misma tabla, donde se incluye su identificador y su clasificación.

V. PROCESOS ETL

Debido a que los datos se encuentran disponibles para aplicaciones estadísticas, los mismos se entregan de forma codificada y agrupados en una única tabla para cada año. Esto hace que sea necesario realizar algunas tareas de procesamiento antes de poblar con datos el esquema relacional implementado, por ejemplo, la limpieza de los datos, su descodificación y la determinación de las relaciones entre los datos. Como parte de una simplificación en la implementación, no se consideró necesaria la creación de *surrogate keys* (llaves autogeneradas) para todas dimensiones, ya que, en el caso de “Región” y “Edad” estos corresponden con valores numéricos.

Además, es importante puntualizar que, inicialmente se consideró la posibilidad de hacer un seguimiento entre períodos para cada persona u hogar. Sin embargo, al analizar los datos se constató que no había evidencia estadística de que las llaves tuvieran correspondencia entre los diferentes períodos, de manera que no sería válido hacer este tipo de análisis. A pesar de esto, en las nuevas encuestas, el INEC está incluyendo un nuevo identificador [14] que posibilita hacer este seguimiento, por lo que para futuras implementaciones si es eficaz hacer esta distinción entre las dimensiones.

A. Extracción

Los datos de la Encuesta Nacional de Hogares se descargaron desde el sitio web del INEC para cada año disponible, lo que comprende el periodo 2010-2016. Los mismos solo se encuentran disponibles en formato *sav*, el cual corresponde con un tipo de archivo de hoja electrónica propietario que requiere del software estadístico *IBM SPSS* para su lectura y procesamiento. Por esta razón, se empleó este programa para procesar cada archivo y exportarlo a *SQL Server* en formato de tabla, lo que corresponde con un almacenamiento temporal que incluye un total de 272,909 registros.

B. Transformación

Las tareas asociadas con la preparación de los datos consisten en efectuar su descodificación de acuerdo con el diccionario provisto por el INEC [14], además de asignar una llave sustituta en los casos requeridos y establecer categorías a los datos que pueden ser agrupados utilizando jerarquías. Un ejemplo de esto es posible observarlo en la figura 4, donde se muestra la construcción de la dimensión "Tipo de actividad". En este proceso se extraen, de forma paralela, los datos desde el almacenamiento temporal asociados con el "Tipo de actividad" que caracteriza a cada persona.

Luego, estos datos son descodificados de acuerdo con la variable que representan, aunque sus categorías deben introducirse de forma manual al sistema, porque el diccionario de datos es un buscador que no ofrece interacción con otras herramientas. En este mismo nivel de tarea, se asigna una llave sustituta que permite identificar tanto la condición como el tipo de actividad asociados. Posteriormente, los datos son ordenados y agrupados, lo que permite asignar los conjuntos correspondientes con el nivel de jerarquía "Condición de actividad". Con ello, los datos están preparados y se procede con su carga a la tabla correspondiente.

En el caso de las mini dimensiones, los datos son generados al unir los posibles valores de cada una de las variables mediante un producto cartesiano usando el operador de *cross join,* lo que crea todas las posibles combinaciones entre sí identificadas con una llave sustituta. En este escenario es importante considerar que parte de las combinaciones de variables no se materializa dentro de los datos recabados, ya que no existen personas u hogares con las características incluidas en algunos registros. Es por esta razón que se considera innecesario que todos estos datos sean cargados para la etapa de análisis, pudiendo suprimir las que no se utilizan. Para este tipo de dimensiones se crearon 645,120 combinaciones para las características del hogar, de las que se utilizaron 8,533, mientras que para los atributos de la persona se establecieron 8,448 registros diferentes, empleándose para el análisis 1,601.

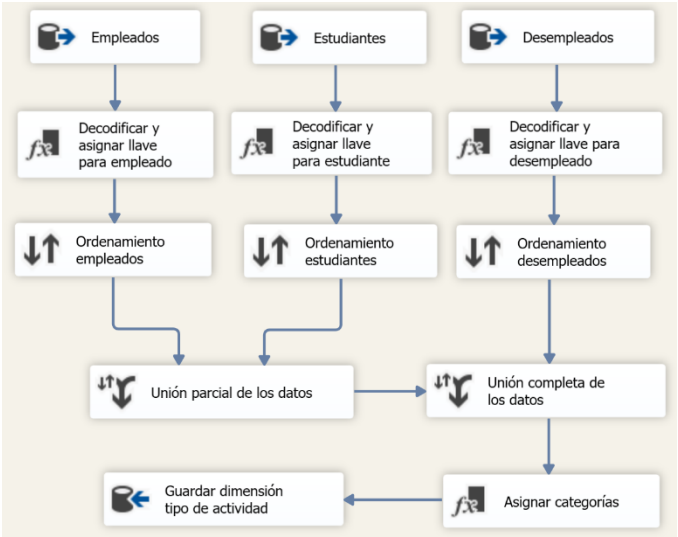

Fig. 4. Ejemplo de la transformación y carga de los datos para la dimensión "Tipo de actividad" utilizando *SQL Server Integration Services*

Para el caso de las tablas de hechos, las transformaciones requeridas implican efectuar la búsqueda de la combinación que agrupa las características correspondientes con cada "Hogar" o "Persona" en la mini-dimensión, como se observa en la parte (a) del ejemplo en la figura 5, que se asocia al proceso del enfoque "Hogar". Además, es necesario asignar las llaves primarias, donde se realiza una concatenación de variables que permite identificar a cada hogar o persona, lo que corresponde con la sección (b) del ejemplo expuesto para el caso del hogar. Con lo anterior, los datos de la tabla de hechos se encuentran preparados para su utilización, de manera que se agrupan y se cargan a la tabla asignada.

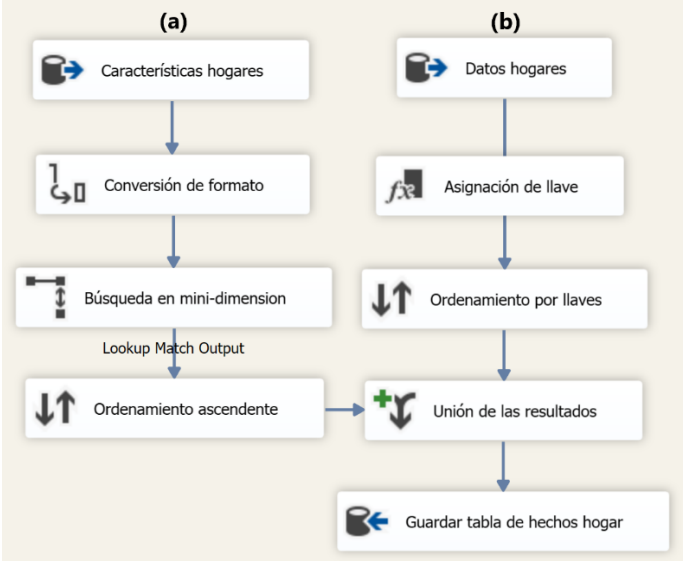

Fig. 5. Ejemplo de la transformación y carga de los datos para la tabla de hechos hogar utilizando *SQL Server Integration Services*. (a) Búsqueda en la mini-dimensión de la llave que agrupa las características del hogar. (b) Carga de los datos del hogar y asignación de su llave primaria.

C. *Carga*

Como se presentó en los dos ejemplos anteriores, luego de especificar las transformaciones necesarias para el procesamiento de los datos, se procede con la carga de los mismos en las tablas definidas en *SQL Server* de acuerdo con el esquema de implementación. Esto se realiza en una ocasión para definir las dimensiones, mientras que, para completar las tablas de hechos, debe repetirse para cada uno de los años de la encuesta, lo que puede realizarse de forma automática al definir una carga por lote.

VI. Implementación del cubo OLAP

Cuando se ha completado la carga de los datos en las tablas del esquema multidimensional se procede, mediante la herramienta *SQL Server Analysis Services*, con la creación de un cubo OLAP. Completado esto, es necesario definir las dimensiones factuales como tales desde las tablas de hechos correspondientes, por ejemplo, las dimensiones factuales "Género" y "Año de encuesta" para poder usarlas como cualquier dimensión. Además, debe indicarse de manera explícita que el vínculo existente entre las tablas persona y hogar ocurre entre dos relaciones factuales, de manera que las medidas sean calculadas de forma correcta en los análisis que los utilicen de forma conjunta. Asimismo, para los análisis que incorporen diferentes períodos, es necesario crear una función de agregación que anualice las medidas que tienen como unidad de tiempo mes o semana, al multiplicarlas por 12 o 52, según sea el caso. También, se requiere que esta función permita que las medidas se sumen al realizar la operación de *roll-up*, es decir, cuando los datos sean agrupados. Con estas observaciones completadas, se implementa el cubo expuesto en la figura 6, mismo que puede ser utilizado en *Microsoft Excel* mediante la opción de tablas dinámicas.

VII. Posibles aplicaciones del modelo

Una de las razones por las que se plantea el desarrollo del sistema propuesto se relaciona con el hecho de considerar que constituye un tipo de herramientas que tiene potencial utilidad para el soporte en la toma de decisiones. Esto considerando que los datos públicos suelen ser poco utilizados, de manera que no proporcionan la información con el suficiente valor para diversas áreas gubernamentales o sectores del ámbito privado.

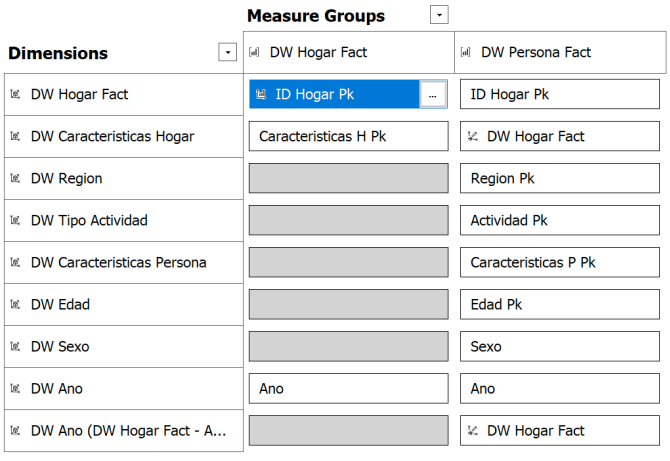

Fig. 6. Vista del uso de dimensiones para el cubo OLAP en *SQL Server Analysis Services*

Por esta razón, se presentan algunos escenarios donde es posible evaluar la funcionalidad de la herramienta descrita anteriormente, así como su potencialidad en la orientación de posibles políticas públicas y decisiones de diferentes grupos sociales. Esto debido a que se posibilita agregar los datos de forma dinámica de acuerdo a como se requiera, simplificando el despliegue de la información de forma tabular, así como la construcción de gráficos que propicien una mejor comprensión de las relaciones existentes entre las variables. Los ejemplos son desarrollados empleando *Microsoft Excel*, ya que corresponde con una herramienta bien conocida por muchos usuarios, aunque, de igual manera, es posible utilizarse en otro tipo de hojas electrónicas, como es el caso de *Libre Office*.

A. *Escenario 1: Comparación de cantidad de horas laboradas por zonas geográficas*

Una forma mediante la cual es posible establecer que tanto trabajo realiza una persona corresponde con la cuantificación de las horas que ha laborado por determinado periodo de tiempo. Esta es una medida de utilidad debido a que permite realizar comparaciones de forma homogénea entre los diferentes tipos de trabajo. Por ejemplo, es complejo determinar qué tan productivo es un programador en comparación con un recogedor de café, ya que desempeñan labores muy diferentes: aun así, es probable que ambos trabajen una cantidad de horas por semana similar, siendo esto lo que define que tanto trabajo realizaron.

Por lo anterior, un posible análisis que se puede realizar con este sistema corresponde con un diagnóstico de la cantidad de tiempo que se labora en Costa Rica, como se puede observar en la figura 7. Esto aunado al hecho de ser una medición que, de acuerdo con las perspectivas del empleo de la OECD, ubica a Costa Rica como el segundo país, entre los miembros y posibles nuevos integrantes de esta organización, con la mayor cantidad de horas laboradas en promedio por año, solo por detrás de México [15].

Para este ejemplo, se visualiza la distribución de la cantidad de trabajo entre las diferentes regiones del país para el año 2013. Además, se incorpora un factor adicional que contempla la migración de personas desde su región de nacimiento hacia otras áreas. Con esto se construye el gráfico de la figura 7, donde cada línea representa la cantidad promedio de horas por semana que laboran las personas de acuerdo con su zona de nacimiento. Asimismo, los segmentos del eje de las abscisas constituyen la región donde residen estos trabajadores.

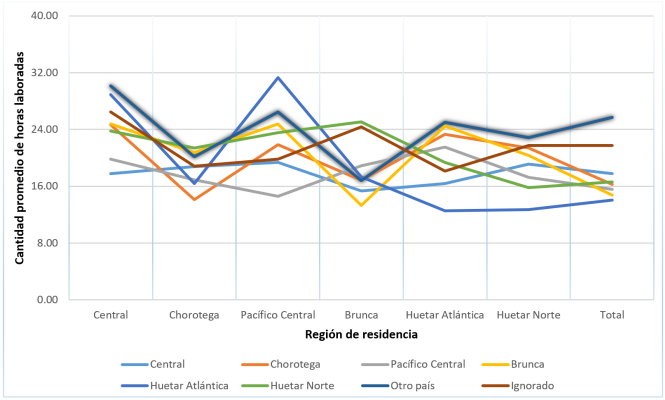

Fig. 7. Cantidad semanal promedio de horas laboradas por zona de residencia y nacimiento en el año 2013

Detallado el gráfico es posible analizar algunas situaciones relevantes como la cantidad de trabajo ofrecido por las personas que han migrado desde otros países, representado mediante la línea azul resaltada del gráfico. En este caso es posible entrever que, para tres de las regiones y el valor total, estos trabajadores laboran en promedio más horas que sus contrapartes costarricenses. Este escenario pone en perspectiva situaciones como la importancia de los migrantes en la fuerza laboral del país, a pesar de las críticas que suele tener la inmigración, por ejemplo, desde Nicaragua. Además, al conocerse las regiones donde se concentra la labor de estos trabajadores, se podrían orientar políticas en aras de promover la formalización de sus trabajos y la correcta recaudación de las cargas sociales.

Adicionalmente, se puede observar que la región Central tiene, en la mayoría de los casos, una cantidad promedio de horas laboradas relativamente más alta que las demás zonas. Esto puede ser un indicador de que las fuentes de empleo se concentran en esta área, siendo desfavorable para las demás al generar subempleo. Asimismo, es posible visualizar que existen diferencias en la cantidad de horas que laboran las personas en una región diferente a la de su origen, por ejemplo, el caso de la región Brunca, lo que podría ser un factor que explique su condición de movilidad al requerir sitios con mayor empleo.

B. Escenario 2: Comparación de cantidad de horas laboradas de las mujeres según diferentes características

De forma análoga, haciendo uso de las mismas medidas que en el escenario anterior, se plantea un caso donde se filtran los datos para obtener una perspectiva diferente de la información. Esto posibilita considerar una situación donde solo se muestren, por ejemplo, la cantidad promedio de horas laboradas por mujeres adultas y adultas jóvenes de acuerdo con su región de nacimiento y su estado civil, con lo que se construye el gráfico de la figura 8. En este caso, es posible observar una situación de interés para todas las zonas, relacionada con la existencia de una menor cantidad de horas laboradas en promedio por aquellas mujeres cuyo estado civil es casada, en unión libre o viuda, en contraste con los demás estados conyugales. Lo anterior puede ser un indicador de que existe una mayor ocupación de las mujeres con estados civiles asociados a la conformación de una familia en actividades externas a la fuerza laboral. Esto comprende labores como la manutención de un hogar y el cuidado de hijos pequeños.

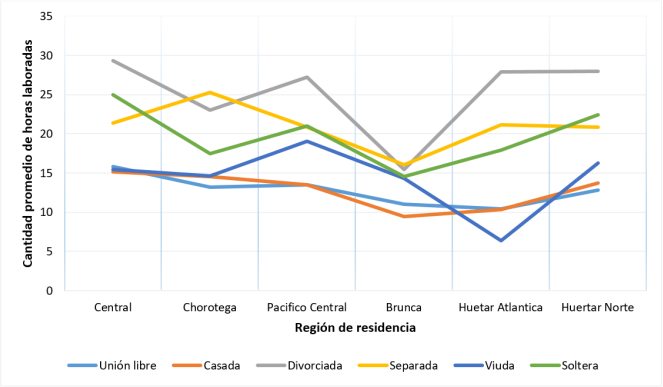

Fig. 8. Cantidad semanal promedio de horas laboradas por mujeres de acuerdo con su región de residencia y estado civil en el año 2013

Asimismo, este tipo de análisis podría ofrecer información valiosa para impulsar la reinserción al mercado laboral de estos grupos, así como un posible control para evaluar que se den condiciones laborales equitativas entre los distintos géneros.

C. Escenario 3: Evaluación de la distribución de programas sociales

Debido a la carencia de herramientas, es usual que en diversas áreas de administración pública resulte complejo determinar el alcance que tienen sus programas de ayuda social. Esto implica que, en múltiples ocasiones, los proyectos que estas entidades realizan no tengan la cobertura que se desea, ya que al ser complejo determinar su rendimiento, pueden resultar beneficiados grupos sociales diferentes a los que se deseaba. Por esta razón, se considera que un posible análisis donde se utilice el modelo multidimensional facilita evaluar el rendimiento de programas sociales, lo que permite conocer cómo están siendo ejecutados y clarificar sobre posibles acciones de mejora.

Un escenario a evaluar corresponde con el análisis de los subsidios que ofrece el gobierno mediante dinero en efectivo. Estos, de acuerdo con los objetivos de las instituciones de ayuda social, deben ser dirigidos a hogares con bajos recursos económicos, por lo que sería esperado que los datos reflejen esta situación. Con esto se plantea un primer escenario donde se contrasta la cantidad mensual promedio de subsidios recibida por los hogares con su respectiva área de construcción para el año 2016, lo que se observa en los diferentes segmentos del gráfico de la figura 9. Además, se incluye una clasificación adicional representada por las distintas barras en cada segmento, que corresponde con las regiones donde se ubican las viviendas. Esto permite entrever que la mayor cantidad de subsidios se concentran en los hogares de menor tamaño, y en la zona Brunca se distribuye más ayuda respecto con las otras regiones.

Con estos datos es posible definir un segundo escenario con un mayor nivel de detalle al incluir la calificación de calidad recibida por los hogares, detallada como buena, regular o mala. Esto se plantea para las viviendas con un tamaño menor a 40 metros cuadrados, ya que corresponden con los segmentos que reciben una mayor cantidad de ayudas para las regiones descritas. Con lo anterior se construye el gráfico de la figura 10, donde se observa, para casi la totalidad de los casos, que la distribución de los subsidios decrece a medida que la calidad de la vivienda aumenta, independientemente de la región donde se encuentre. Esto permite notar que los subsidios son entregados a aquellas familias que residen en hogares con mayor deterioro, siendo estos los que más requieren subsidios del Estado.

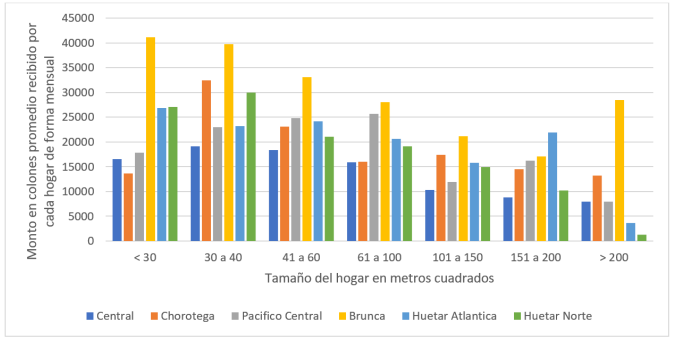

Fig. 9. Distribución promedio de los subsidios por tamaño del hogar y ubicación para el año 2016

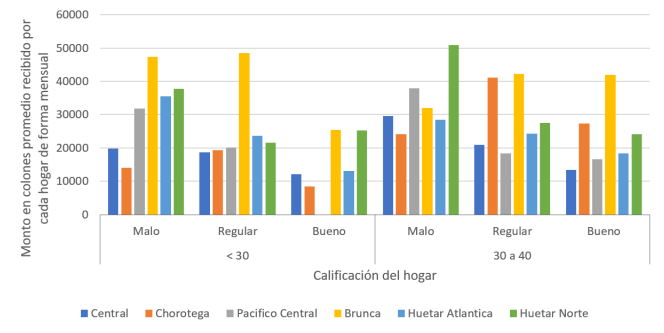

Fig. 10. Distribución promedio de los subsidios por calificación del hogar y ubicación, para viviendas con un tamaño menor a 40 metros cuadrados en el año 2016

De acuerdo con los dos escenarios planteados en las figuras 9 y 10, es posible entrever como los subsidios se distribuyen de manera acorde con lo que se esperaría para este tipo programas de ayuda social: que sean dirigidos a las familias de bajos recursos. Esto se manifiesta al percibir que las viviendas de mayor tamaño reciben menos subsidios, además de que, al explorar con un mayor nivel de detalle, los hogares con una calificación mala reciben más ayudas. Adicionalmente, al definir las divisiones por zonas, es destacable observar que la mayoría de los subsidios se distribuyen en las regiones Brunca y Chorotega, las cuales corresponden con áreas lejanas al centro del país y que históricamente poseen mayores niveles de pobreza. Por lo anterior, de acuerdo con los escenarios planteados, es posible observar que mediante la utilización de un modelo multidimensional se pueden contrastar diferentes enfoques que permitan medir el alcance de programas gubernamentales de este tipo.

D. Escenario 4: Valoración de la distribución de las bonificaciones de acuerdo con diferentes factores

Un tema ampliamente discutido en diferentes ámbitos académicos y sociales corresponde con la existencia de una brecha importante respecto a los ingresos que perciben las personas de acuerdo con su género. Esto suele representar un problema social, ya que pone en evidencia patrones culturales que socavan los diferentes esfuerzos por promover la equidad de género y la igualdad de condiciones laborales independientemente del sexo. Es por esta razón que se plantea un posible análisis donde se manifieste la forma en cómo se distribuyen los ingresos por género de acuerdo con el nivel de estudios alcanzado por las personas y las horas que trabajan semanalmente. Estas son dos características con las que se procura tener un enfoque objetivo sobre la igualdad de condiciones en el trabajo. Además, no se utiliza el ingreso principal de las personas, como puede ser su salario, sino las bonificaciones recibidas en promedio de forma mensual. Lo anterior, debido a que estas suelen otorgarse con mayor discreción, de manera que es donde pueden generarse mayores disparidades.

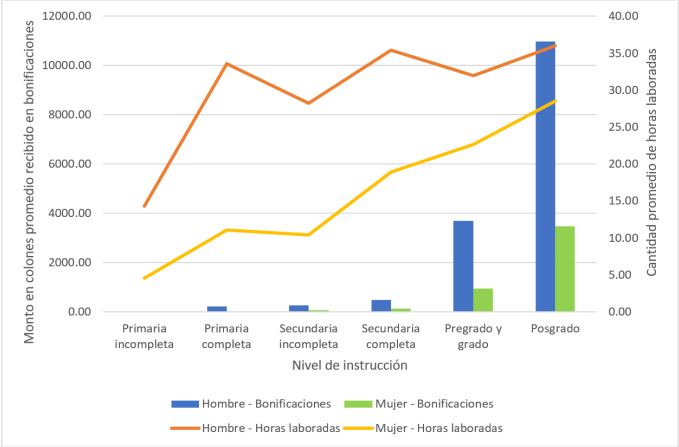

Fig. 11. Distribución promedio de las bonificaciones mensuales y cantidad semanal promedio de horas laboradas por nivel de instrucción y género en el año 2016.

En la figura 11 se presentan dos gráficos superpuestos donde es posible observar, mediante barras, la forma en cómo se distribuyen las bonificaciones de acuerdo con el nivel de instrucción de las personas y su género. Asimismo, utilizando líneas y el eje vertical secundario, se representa la cantidad de horas laboradas en promedio por semana para cada género acorde con los diferentes niveles de educación. Con esta disposición de los datos, es posible observar que las bonificaciones se otorgan, principalmente, a quienes cuentan con estudios universitarios y laboran, en promedio, una mayor cantidad de horas. Asimismo, es posible notar una disparidad significativa respecto a la distribución de este tipo de ingresos al hacer la distinción por el género, de manera que los hombres reciben más bonificaciones que las mujeres. Tal hecho ocurre a pesar de que los datos corresponden con los mismos niveles de instrucción y la diferencia entre la cantidad de horas laboradas entre sexos no es tan marcada. Esto permite entrever que las diferencias respecto a la distribución por género de las rentas existen en un país como Costa Rica y es necesario aunar esfuerzos que posibiliten la reducción de esta brecha.

E. Escenario 5: Influencia del bono de vivienda y la tenencia de la propiedad en la cantidad de artículos con los que cuenta el hogar

Una manera con la cual se puede medir la influencia que tienen programas de apoyo social como el bono de construcción de la vivienda sobre el nivel de gastos de un hogar corresponden a una comparación cruzada con respecto a las residencias que no reciben este tipo de ayuda. Además, al incluir la tenencia en propiedad de la construcción, por ejemplo, si la misma está siendo pagada o es alquilada, es posible determinar si existe un gasto familiar importante para cubrir este rubro. Con esta información es posible construir el gráfico de la figura 12, donde se muestra, mediante barras, la cantidad de celulares, computadoras y automóviles con los que cuenta el hogar, lo que se realiza en escala logarítmica para facilitar su comparación. Adicionalmente, se incluye el ingreso promedio de las personas que residen en la vivienda, representado con una línea y el eje vertical secundario, ya que es una medida importante para cuantificar el poder de adquisición.

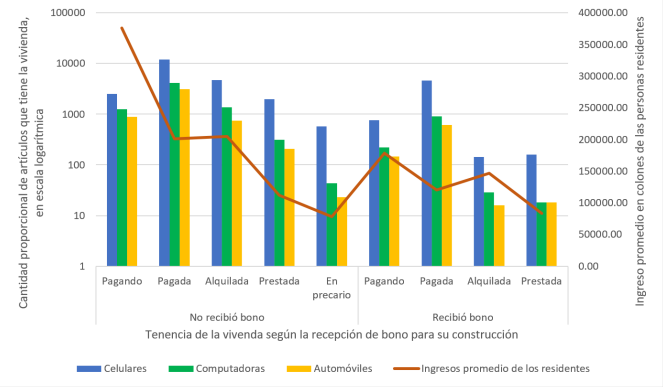

Fig. 12. Cantidad de artículos con los cuenta una vivienda e ingreso según la recepción del bono y la tenencia de la propiedad para el año 2015.

Con lo anterior, es posible observar que, en todos los casos, es de gran relevancia la posesión de teléfonos celulares, además de que la cantidad de artículos en el hogar es mayor cuando la vivienda se encuentra en propiedad, es decir, que no implica un gasto periódico. De un modo similar, se puede observar que quienes se encuentran pagando la residencia suelen poseer menor cantidad de artículos, a pesar de tener un mayor ingreso, lo que se puede explicar por el egreso asociado a la cancelación de la vivienda, ya sea adquirida con un bono o no. Asimismo, es posible notar que las familias que reciben esta ayuda social suelen poseer relativamente menores ingresos y menos artículos en el hogar. Esto permite establecer que los bonos están siendo distribuidos de forma correcta, al estar enfocados en los hogares que lo requieren dada su condición económica.

VIII. Conclusiones

Como una síntesis final de este trabajo, es posible definir los siguientes puntos como conclusiones del mismo:

- La implementación realizada del modelo multidimensional opera de manera correcta para un enfoque no transaccional como lo son las encuestas. Esto posibilita que su adopción en bases de datos públicas permita ofrecer herramientas para el análisis de datos que sean sencillas y que operen en un ambiente relativamente conocido, como lo es la hoja electrónica.

- El potencial de análisis que ofrecen las herramientas OLAP es bastante alto para su uso con datos que permitan una mejor toma de decisiones en sectores como el gubernamental, así como la definición de controles que permitan una mejor administración de los proyectos públicos. Esto debido a que suelen ser áreas con poca orientación hacia la utilización de datos, de manera que es necesario el desarrollo de aplicaciones de este tipo para faciliten la adopción en el futuro de herramientas con mayor potencialidad de análisis.

- La disponibilidad de una gran cantidad de datos públicos y abiertos posibilita la implementación de modelos multidimensionales que permitan facilitar su acceso y soportar la toma de decisiones en múltiples ámbitos. Esto faculta que una adopción de este tipo de herramientas incremente la utilización de datos en el sector público y privado, sin existir una necesidad expresa para crear nuevos procesos generadores de información. Además, este es un modelo que puede utilizarse en el desarrollo de futuros trabajos donde se utilicen datos adicionales que permitan ampliar los escenarios de análisis y las áreas en las que puede ser aplicado.

Referencias

[1] Instituto Nacional de Estadística y Censos, "Encuesta Nacional de Hogares," Instituto Nacional de Estadística y Censos, 16 Diciembre 2016. [En línea]. Disponible: http://www.inec.go.cr/encuestas/encuesta-nacional-de-hogares. [Accedido 15 Marzo 2017].

[2] C. Howson and A. Duncan, "ITScore Overview for BI and Analytics," Gartner, 24 Septiembre 2015. [En línea]. Disponible: https://www.gartner.com/doc/3136418/itscore-overview-bi-analytics. [Accedido 24 Marzo 2017].

[3] Global Working Group on Big Data, "Using Big Data for the Sustainable Development Goals," United Nations Statistical Commission, Diciembre 2015. [En línea]. Disponible: https://unstats.un.org/bigdata/taskteams/sdgs/. [Accedido 17 Marzo 2017].

[4] Ministerio de Planificación Nacional y Política Económica, "Costa Rica 2030: Objetivos de Desarrollo Nacional," Octubre 2013. [En línea]. Disponible: https://www.mideplan.go.cr/prensa/142-noticias-anteriores/1250-mideplan-le-invita. [Accedido 16 Marzo 2017].

[5] Observatorio del Desarrollo, "Sitio web principal," Universidad de Costa Rica, 2016. [En línea]. Disponible: http://odd.ucr.ac.cr/. [Accedido 15 Marzo 2017].

[6] A. Gómez, "Tendencias del desarrollo costarricense," Universidad de Costa Rica, 2013. [En línea]. Disponible: http://odd.ucr.ac.cr/proyectos/tendencias-del-desarrollo-costarricense. [Accedido 20 Marzo 2017].

[7] Observatorio del Desarrollo, "Costa Rica en cifras," Universidad de Costa Rica, 2013. [En línea]. Disponible: http://www.odd.ucr.ac.cr/proyectos/costa-rica-en-cifras/. [Accedido 16 Marzo 2017].

[8] Centro Centroamericano de Población, "Inicio," Universidad de Costa Rica, 2017. [En línea]. Disponible: http://ccp.ucr.ac.cr/. [Accedido 17 Marzo 2017].

[9] Centro Centroamericano de Población, "Sistema de Consulta a Base de Datos Estadísticos," Universidad de Costa Rica, 2012. [En línea]. Disponible: http://censos.ccp.ucr.ac.cr/index.php/usuarios_c/index. [Accedido 16 Marzo 2017].

[10] Centro Centroamericano de Población, "Cómo hacer consultas con el PDQ-Explore," Universidad de Costa Rica, 2012. [En línea]. Disponible: http://consultas.ccp.ucr.ac.cr/ayuda.html. [Accedido 26 Marzo 2017].

[11] G. Argüello, "Las encuestas de hogares en América Latina: Estado del situación y perspectiva," Instituto Nacional de Estadística y Censos, Octubre 2015. [En línea]. Disponible: http://www.cepal.org/sites/default/files/events/files/2015-10-tallereh-c-giselle-arguello.pdf. [Accedido 16 Marzo 2017].

[12] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed., Indianapolis, IN: John Wiley & Sons, 2013.

[13] E. Malinowski and E. Zimányi, Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications, 2nd ed., Berlin, Heidelberg: Springer-Verlag, 2009.

[14] Instituto Nacional de Estadística y Censos, "Diccionario de Datos - Encuesta Nacional de Hogares 2016," Instituto Nacional de Estadística y Censos, 16 Diciembre 2016. [En línea]. Disponible: http://sistemas.inec.cr/pad4/index.php/catalog/165/data_dictionary. [Accedido 26 Marzo 2017].

[15] Organisation for Economic Co-operation and Development, "OECD Employment Outlook 2016," OECD Publishing, 7 Julio 2016. [En línea]. Disponible: http://dx.doi.org/10.1787/empl_outlook-2016-en. [Accedido 5 Abril 2017].