Capítulo 18

Estimación de la pobreza utilizando datos de teléfonos celulares: evidencia de Guatemala

Marco Antonio Hernández Ore, Lingzi Hong, Vanessa Frías-Martínez, Andrew Whitby y Enrique Frías-Martínez

Extracto

La dramática expansión del uso de teléfonos móviles en países en desarrollo ha resultado en un incremento de fuentes ricas y mayormente intactas de información sobre las características de las comunidades y regiones. Los Registros de Detalles de Llamadas (CDR) obtenidos de los teléfonos celulares proveen una información altamente granular en tiempo real que puede ser usada para evaluar el comportamiento socioeconómico incluyendo consumo, movilidad y patrones sociales. Esta nota examina los resultados de un análisis CDR enfocado en cinco departamentos administrativos en la región suroeste de Guatemala, el cual usó datos de teléfonos celulares para predecir los índices de pobreza observados. Sus descubrimientos indican que los métodos de investigación basados en CDR tienen el potencial para replicar las estimaciones de pobreza obtenidos de las formas tradicionales de recolección de datos, como encuestas en hogares o censos, por una fracción del costo. En particular, los CDR fueron más útiles en predecir la pobreza urbana y total en Guatemala con más precisión que la pobreza rural. Además, mientras que los estimados de pobreza producidos por los análisis CDR no encajan perfectamente en aquellos generados por encuestas y censos, los resultados muestran que obtener más información exhaustiva podría mejorar enormemente su poder predictivo. El análisis CDR tiene especialmente aplicaciones prometedoras en Guatemala y otros países en desarrollo, lo cuales sufren altos índices de pobreza e inequidad, y donde los limitados recursos presupuestarios y fiscales complicarían la tarea de recolección de datos. Además, destacan la importancia de focalizar con precisión los gastos públicos para lograr su máximo impacto antipobreza.

Introducción

El explosivo crecimiento de las redes de telecomunicaciones en los países en desarrollo está arrojando una riqueza sin precedentes de datos altamente

granular en tiempo real, y los gobiernos solo han comenzado a aprovechar el enorme potencial de esta nueva fuente de información. Esta nota explora las metodologías analíticas para usar datos de teléfonos celulares agregados y cifrados para mapear la distribución de la pobreza en Guatemala. Para estimar los índices de pobreza, Guatemala, como muchos otros países en desarrollo, depende de la conducción y análisis de encuestas en hogares y censos de población que son tanto costosos como exigentes administrativamente. Por contraste, el análisis basado en datos de teléfonos celulares en combinación con el aprendizaje máquina tiene el potencial de generar información confiable y oportuna sobre la distribución espacial de la pobreza en hogares a un costo mucho menor que las tradicionales encuestas de hogares o censos de población.

Pobreza en Guatemala

Los índices de pobreza en Guatemala son mayores que en otros países de ingresos medios comparables, y la distribución de la pobreza refleja una serie de dimensiones regionales, rural/urbana y étnicas superpuestas. Contrario a la tendencia observada en otros países de Latinoamérica, los índices de pobreza en Guatemala han crecido en años recientes. Los índices de pobreza[1] crecieron desde el 55% en 2001 al 60% en 2014, mientras que la cantidad de personas que viven por debajo de la línea de pobreza se incrementó en unos 2,8 millones. Los índices de pobreza varían dramáticamente por departamento administrativo. En 2014, el departamento más pobre de Guatemala, Alta Verapaz, tuvo un índice de pobreza del 83% y un índice de pobreza extrema[2] del 54%. Mientras tanto, los índices de pobreza y pobreza extrema en el departamento más rico, donde se encuentra localizada la Ciudad de Guatemala, fueron mucho más bajos ubicados en el 33% y 5%, respectivamente. Además, mientras que las áreas urbanas son ahora el hogar de una mayoría de la pobreza del país, los índices de pobreza se mantienen sustancialmente mayores en áreas rurales. En 2014, el 35% de la población rural estaba viviendo en pobreza extrema, comparado con el 11% de la población urbana. Los índices de pobreza son también significativamente mayores entre la población indígena de Guatemala. La gente indígena representa el 42% de la población total del país, pero en 2014 representaba el 52% de los pobres y el 66% de los extremadamente pobres en el país[3].

1. Los índices de pobreza moderada reflejan el consumo familiar per cápita equivalente a $4.00 al día en términos de poder adquisitivo.
2. Los índices de pobreza extrema reflejan el consumo familiar per cápita equivalente a $2.50 al día en términos de poder adquisitivo.
3. Sanchez, Scott y Lopez (2016).

Los altos índices de pobreza en Guatemala y la persistente inequidad en ingresos son reflejados en los débiles indicadores del desarrollo humano en el país. El acceso limitado y desigual a los servicios públicos como educación y el cuidado de la salud han restringido la formación del capital humano. Mientras que una falta de infraestructura básica ha incrementado los costos de producción y transporte y ha reducido las oportunidades de trabajo disponibles para los pobres. En conjunto, estos factores están contribuyendo al declive a largo plazo de la productividad económica. Mientras tanto, los bajos ingresos de impuestos en Guatemala limitan su capacidad para una política de redistribución fiscal y gastos en pro de los pobres. Los altos índices de pobreza y las fuertes restricciones fiscales de Guatemala destacan la crítica importancia de focalizar efectivamente el gasto público.

El rol de la información en la reducción de la pobreza

Las estrategias efectivas para reducir la pobreza requieren información detallada sobre la actual distribución geográfica de la pobreza y las características de los hogares que viven debajo de la línea de pobreza. Los censos y las encuestas de hogares pueden dar luz a un amplio rango de indicadores económicos y sociales. Sin embargo, estos métodos son costosos y consumen tiempo e implementarlos requiere de capacidad institucional. Además, las condiciones locales adversas como conflictos violentos, altos índices de crímenes o inestabilidad política puede hacer que las encuestas personales sean imposibles en ciertas áreas. Como resultado, los políticos responsables deben con frecuencia basar sus decisiones críticas en información incompleta u obsoleta.

Los científicos sociales están usando crecientemente el análisis de *big data* para suplementar más fuentes tradicionales de información. Las imágenes de satélites, registros de sensores (es decir, tráfico, clima), aplicaciones de teléfonos inteligentes e información de teléfonos celulares —el tema de esta nota— ya han arrojado perspectivas importantes en numerosos campos. A diferencia de las encuestas en hogares, las cuales están específicamente diseñadas para abordar ciertas preguntas de investigación, los grandes conjuntos de datos son usualmente recolectados en un contexto no investigativo, usualmente como el derivado de una actividad comercial o servicio público. Analizar *big data* requiere de nuevos métodos de investigación, muchos de los cuales están todavía en etapas iniciales de su desarrollo. Las metodologías de investigación emergentes basadas en Registros de Detalles de Llamadas (CDR) y técnicas avanzadas de aprendizaje máquina o *machine learning* tienen aplicaciones

especialmente prometedoras en países en desarrollo, ya que ellas pueden potencialmente generar datos confiables de pobreza a un costo mucho menor que las encuestas de hogares convencionales.

El análisis CDR puede jugar un rol vital al llenar los huecos espaciales y temporales dejados por los métodos tradicionales de investigación. Al hacer inferencias basadas en el uso de redes celulares, el análisis CDR puede proyectar confiablemente la evolución de las dinámicas de pobreza en un marco de tiempo específico. A diferencia de los censos y las encuestas de hogares, el análisis CDR es rápido y relativamente económico y puede ser realizado por un grupo pequeño de estadísticos usando registros que ya fueron recolectados por Operadoras de Redes Móviles (MNO).

Guatemala ofrece un claro ejemplo de los límites de la recolección tradicional de datos. El Censo de Hogares y Población más reciente tiene fecha de 2002 y todos los datos de pobreza nacional están derivados de solo 4 encuestas de hogares realizadas en los últimos 25 años. Por ejemplo, la más reciente encuesta de hogares de 2014 (Encuesta Nacional de Condiciones de Vida, ENCOVI) cubre alrededor de 11.500 hogares, se tardó dos años en completarla a un costo de alrededor de 2 millones de dólares. Por contraste, el análisis CDR realizado para este reporte tuvo un valor de alrededor de 100.000 dólares, y la mayoría de los gastos fue para al desarrollo del algoritmo de computadora, el cual es un costo fijo. Esto, si el análisis CDR es conducido nuevamente con nuevos datos, sería significativamente más económico.

Mientras que este fue el primer análisis de su tipo realizado en Guatemala y diseñado para probar principalmente la validez de varias metodologías, un ejercicio más minucioso requeriría solo de una pequeña fracción del tiempo y recursos humanos relativos a un censo o encuesta tradicional. Además, estos costos serían probablemente menor en iteraciones posteriores del análisis CDR, mientras la innovación y las pruebas tecnológicas son reemplazadas por la implementación rutinaria de técnicas establecidas. Mientras que el análisis CDR no puede reemplazar totalmente los métodos convencionales de investigación, pueden mejorar enormemente su valor al proveer actualizaciones de alta frecuencia e información complementaria. Además, si el análisis CDR puede demostrarse que puede proveer inferencias suficientemente precisas para permitir a los países extender ligeramente el tiempo entre las encuestas tradicionales, podría potencialmente generar un ahorro neto para el presupuesto nacional de investigación.

Antecedentes

Registro de detalles de llamada (CDR)

Las operadoras de redes móviles registran y almacenan datos sobre el uso de los teléfonos de sus clientes, primariamente para propósitos de cobro. Adicionalmente al registrar el consumo de datos celulares, las MNO recolectan información de cada llamada y mensaje. Los datos almacenados no reflejan generalmente el contenido de una llamada o mensaje. En vez de eso, se registran detalles circunstanciales, como hora y duración de la llamada, el tamaño del mensaje, las identidades de las partes involucradas y su información de la red. Se refieren a estos datos en la industria de telecomunicaciones como la CDR.

Figura 1. Ejemplo de registros de detalles de llamadas

Interacción	Dirección	ID Correspondiente	Hora y fecha	Duración de la llamada	ID de la antena
Llamada	Entrada	8f8ad28de134	2012-05-20 20:30:37	137	13084
Llamada	Salida	fe01d67aeccd	2012-05-20 20:31:42	542	13084
Texto	Entrada	c8f538f1ccb2	2012-05-20 21:10:31		13087

Fuente: http://bandicoot.mit.edu/docs/quickstart.html

Adicionalmente a las CDR, las MNO usualmente almacenan ciertos detalles personales sobre sus usuarios, incluyendo sus nombres y dirección de hogar, y en algunos casos su género, edad u otras características. Para clientes de prepago, los cuales son muy comunes en países de ingresos bajos y medios, las MNO típicamente guardan un registro de las recargas de créditos o "incremento de saldo".

Usar CDR en investigaciones sociales y económicas

Aunque las CDR pueden aparentar una serie de datos técnicos y estrechos debido a la dramática expansión del uso del teléfono móvil en las últimas décadas, estos registros pueden proveer una rica fuente de información sobre el comportamiento humano y las características de comunidades. Las CDR pueden ser usadas para inferir ciertos atributos personales sobre un usuario de celular, tales como la ubicación de su hogar. De allí, ellos pueden ser usados para analizar las redes sociales, ya que cada llamada puede ser vista como un vínculo entre los

clientes de un MNO. Este enfoque permite a los investigadores trazar las interacciones sociales, identificar puntos de nexo de las comunidades y examinar cómo se transmite la información a través de grupos y regiones[4].

Caja 1. La geografía virtual de los teléfonos celulares: determinar la ubicación de los usuarios desde los CDR

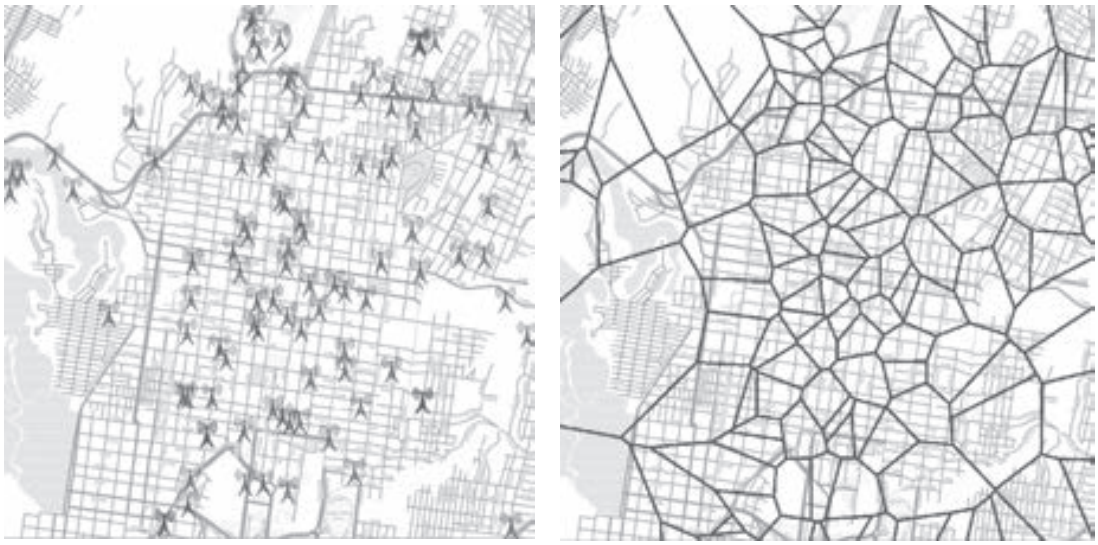
A diferencia de los celulares de satélite, los teléfonos celulares dependen de una red de torres conocidas como estaciones base, las cuales operan dentro de un rango limitado. Esto divide el área de cobertura de la red en "células" individuales. Mientras el cliente se mueve, su conexión a la red es transferida de una torre a la siguiente. Cuando un cliente realiza una llamada, envía un texto o inicia una sesión de datos, la identidad de la torre relevante es registrada en el CDR. Las MNO mantienen una lista de las coordenadas de cada torre, haciendo posible determinar la ubicación general de un teléfono cada vez que es usado.

Mientras que la ubicación precisa del usuario no puede ser identificada, áreas geográficas limitadas llamadas "polígonos de Voronoi" pueden ser construidas basadas en la asunción de que un móvil siempre se conecta a la torre celular más cercana. En realidad, factores como la diferente potencia de las antenas, restricciones de capacidad y terreno puede causar que un móvil se conecte a una antena más lejana. Sin embargo, los polígonos Voronoi continúan siendo una herramienta útil para aproximar la ubicación de un usuario de celular.

La estructura de la red determina el tamaño de cada polígono Voronoi. Ya que tanto el equipo físico de la antena y el espectro móvil tienen capacidad limitada, las MNO tienden a colocar más antenas en áreas de mayor uso para maximizar el rendimiento. Por ende, el tamaño de los polígonos tiende a correlacionarse inversamente con la densidad de población –es decir, los lugares densamente poblados tienden a tener más antenas y polígonos más pequeños–. Sin embargo, algunos lugares con una población residencial pequeña, pero con altos índices de actividad comercial, tales como distritos de negocios, centros comerciales y aeropuertos, pueden tener un número mayor de antenas. Los polígonos pueden variar en diámetro desde unos cientos metros a decenas de kilómetros, dependiendo de la red.

4. Estas aplicaciones son descritas con mayor detalle en Blondel *et al*. (2015).

Figura 2. Muestra de Las torres celulares alrededor de la Plaza Constitución de la Ciudad de Guatemala (izquierda) y los polígonos Voronoi que generan (derecha)



Nota: el Proyecto OpenCellID recolecta ubicaciones de redes celulares basado en los reportes de usuarios voluntarios quienes instalaron una aplicación de participación. Por lo tanto, estas ubicaciones de las torres son aproximadas. Ninguna información oficial de torres celulares ha sido usada para estos gráficos.
Fuente: opencellid.org

Estimar la pobreza en Guatemala usando datos de teléfonos celulares

Esta sección describe el resultado de un estudio reciente de métodos de investigación basados en los CDR en Guatemala que fueron diseñados para evaluar el valor potencial del análisis CDR como una herramienta de investigación socioeconómica. El objetivo del estudio era crear un modelo usando datos de CDR que pudiera predecir con precisión la incidencia observada de pobreza extrema. El estudio se enfocó en cinco municipios en los departamentos administrativos de Quetzaltenango, Suchitepéquez, Sololá, Totonicapán y San Marcos. Juntos, estos departamentos representan el 20% de la población guatemalteca.

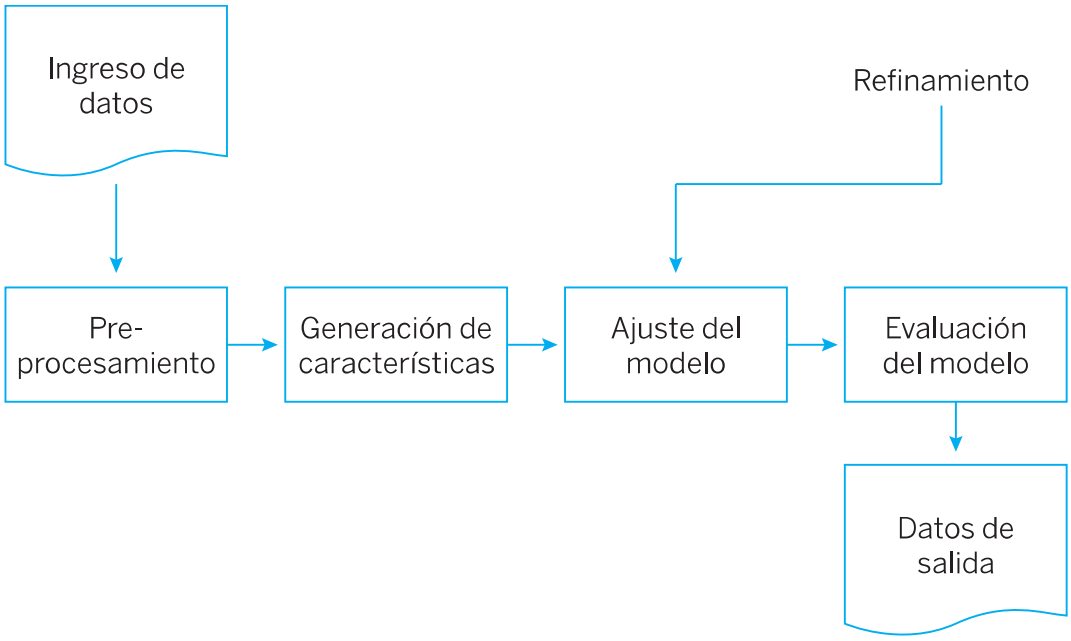
El estudio abordaba tres preguntas:

- ¿Pueden los datos de CDR ser usados para estimar confiablemente los índices de pobreza en Guatemala?

- ¿Son estos estimados más precisos en áreas urbanas, áreas rurales o a nivel nacional?
- ¿Pueden los datos de pobreza derivados de los CDR en el 2006 ser usados para predecir los índices de pobreza en el 2011?

Para responder a estas preguntas, el estudio empleó un enfoque de aprendizaje máquina, el cual es un método altamente general, iterativo para descubrir la relación entre datos de entrada y datos de salida, los cuales en este caso son los registros de teléfonos celulares y los índices de pobreza, respectivamente. La figura número 3 ilustra la metodología de aprendizaje de máquinas.

Figura 3. Pasos típicos en un análisis de aprendizaje de máquinas



Fuente: adaptado de Hong y Frías-Martínez [2015a].

Fuentes de datos

Para poder probar la validez del análisis CDR, sus descubrimientos fueron comparados con los estimados de pobreza del Banco Mundial, los cuales están basados en la Encuesta Nacional de Condiciones de Vida de Guatemala (ENCOVI) para 2006 y 2011 y el Censo de Hogares y Población de 2002. El ENCOVI es una encuesta de hogares tradicionales y su tamaño de muestra no provee estimados confiables a nivel municipal. Sin embargo, las técnicas de estimación de áreas pequeñas[5] que combinan los datos de ENCOVI con los datos del censo permiten

5. Ver Elbers, Lanjouw y Lanjouw (2003).

que la pobreza sea estimada a nivel municipal. Para los propósitos del estudio, estos estimados fueron tratados como "datos reales del terreno". En un análisis de aprendizaje de máquinas, los datos reales del terreno son obtenidos por observación directa, en vez de por un modelaje o inferencia. En este contexto, sin embargo, el término se refiere a los índices de pobreza determinados por métodos de estimación estadísticos estándar, los cuales proveen la única medida existente de la realidad del terreno. Sin embargo, debe tenerse en cuenta que todas las metodologías de estimación de índice de pobreza son predicados en asunciones, y en esta área ningún dato del terreno puede ofrecer una representación perfecta de la realidad.

Los modelos supervisados de aprendizaje máquina, como se describen aquí, requieren un serie de datos de entrenamiento, los cuales comprenden datos de referencia que representan la realidad del terreno. Ya que los CDR se relacionan directamente con las personas, pueden ser considerados datos de registro de unidad. Entonces es el detalle de los datos de la realidad del terreno los que determinan enormemente la resolución del modelo. Esta es usualmente la única opción disponible cuando los datos de la realidad del terreno están basados en estimaciones usando datos de encuestas de hogares, en los cuales ningún número de teléfono celular es recolectado durante la encuesta. En este caso, las características de nivel individual son extraídas de los CDR y luego combinadas para formar agregados estadísticos en el nivel geográfico elegido (es decir, medio, mediano, máximo o cuantiles por región). Un tamaño de muestra relativamente pequeño (por ejemplo, los 338 municipios de Guatemala) significa que la validación interna, como la validación en cruce o pruebas ocultas de datos, es difícil, por lo que pueden ser requeridas algunas validaciones externas.

Los índices de pobreza fueron calculados para cada municipio. Los índices agregados de pobreza rural, urbana y general estuvieron disponibles para 2006, pero solo los índices de pobreza rural estuvieron disponibles para 2011[6]. El estudio uso datos CDR agregados y cifrados para agosto de 2013, el cual se superpone con el periodo de encuesta del ENCOVI. En 2013 Guatemala tenía 140 cuentas celulares por cada 100 personas[7]. El modelo probó dos tipos de predicciones: (1) predicción de la misma encuesta, tal como predecir los índices de pobreza urbana de 2006 basado en un modelo de relación entre los datos de ENCOVI de 2006 y los CDR de 2013; y (2) predicción de encuestas diferentes, tal como

6. Estos datos provienen de un censo de las áreas rurales de 2011 diseñado para recolectar información para programas de protección social. Ningún censo nacional se realizó ese año.
7. Indicadores de Desarrollo Mundial 2016. Esto refleja múltiples cuentas por persona. Mientras que esto no indica que cada persona tiene un teléfono celular, sugiere que el uso de teléfono celular es alto.

predecir los índices de pobreza rural de 2011 basado en un modelo de la relación entre los datos de ENCOVI de 2006 y los CDR de 2013. La predicción de encuestas mismas es más similar a la aplicación en el mundo real del modelo de datos celulares conocidos como "relleno espacial" o "extrapolación espacial", mientras que la predicción de las encuestas diferentes es un ejemplo de "extrapolación de tiempo."

Preprocesamiento

Limpieza de datos y enriquecimiento

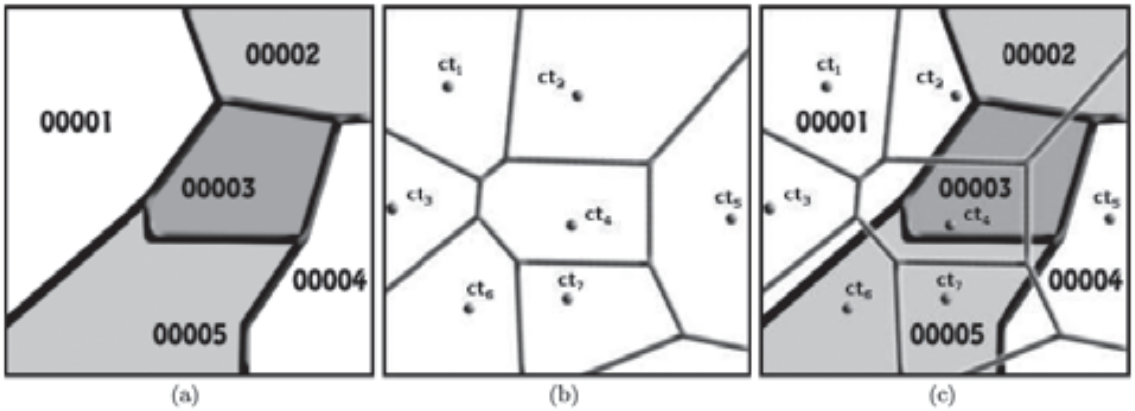
Los datos crudos de CDR son invariablemente ruidosos y requieren de un preprocesamiento antes de poder ser analizados. Las fuentes primarias de ruido en datos son: (1) huecos o inconsistencias causadas por decisiones de operación de las redes de las MNO, incluyendo cambios tecnológicos que afectan la comparabilidad de los CDR; y (2) la presencia de líneas de negocios, números para recargar textos u otras conexiones que no reflejan comunicaciones entre clientes individuales. El preprocesamiento comienza al identificar los CDR inconsistentes o irrelevantes y eliminarlos de la serie de datos. Cada cliente en la serie de datos de CDR es entonces asignado a una ubicación de hogar, la cual es generalmente inferida basada en la red celular en la cual ese cliente se encuentra más activo después de las 6pm. Sin embargo, si esa célula tiene 30% más actividad que la célula más activa siguiente durante el mismo periodo, se asume que el usuario es itinerante y no se asigna una ubicación de hogar.

Armonización espacial

Asegurar que todas las series de datos compartan una escala espacial común es un paso importante en la preparación de datos. Mientras que la red celular es la unidad espacial natural para los datos de CDR, los índices de pobreza son generalmente calculados usando límites administrativos, en este caso el *municipio*. Estas estructuras espaciales deben ser reconciliadas para poder analizar los datos.

En el modelo presentado a continuación, los datos de pobreza están trazados sobre las redes celulares. A cada célula le es asignado un promedio de área pesada de los índices de pobreza en los municipios que cubre. Las redes celulares ubicadas enteramente dentro de un municipio son asignadas al valor de dicho municipio. Este proceso, ilustrado en la figura 4, asegura que todas las series de datos usen la geografía común de la red celular, la cual entonces se convierte en la unidad de análisis. Los enfoques más complejos podrían ser adoptados, tales como ponderar la distribución por densidad de población.

Figura 4. Un ejemplo de armonización espacial



Nota: Unidades Geográficas Unidas (a) de los datos de encuestas (b) con la geografía de las redes celulares definidas como Diagramas Voronoi. (c) La incidencia imputada de pobreza para el polígono 0003 será "x" por ciento de aquel en ct2, "y" por ciento en ct4 y "z" por ciento en ct5. Donde x+y+z =100, representando el área completa del polígono 00003.
Fuente: reproducido de Frías-Martínez *et al.* (2012), figura 1.

Generación de características

Bajo el enfoque de aprendizaje de máquinas la "generación de características es el proceso de colapsar una rica serie de datos multidimensionales hacia un número pequeño de dimensiones cuidadosamente elegidas, las cuales son estadísticas de los datos subyacentes. Estas características entonces forman una representación intermedia de los datos y son la base para el modelo final basado en regresión o basado en clasificación. Por ejemplo, el reconocimiento de la letra escrita a mano es una tarea clásica de aprendizaje de máquinas en la cual una imagen de un dígito escrito a mano representa una serie de datos altamente multidimensional. En algunos casos, cada imagen puede ser 64 por 64 píxeles, para un total de 4.096 dimensiones. Ya que esto puede ser demasiado para que un modelo de clasificación funcione bien, un número más pequeño de características son extraídas usando algoritmos preexistentes bien conocidos. Estas características usualmente representan la presencia o ausencia de características geométricas de un nivel mayor como filos, curvas y esquinas. Este simplifica el modelo final, acelerando el proceso de entrenamiento[8].

8. El paradigma "característica" en el aprendizaje de máquinas está gradualmente siendo superado por enfoques de "aprendizaje profundo", los cuales trabajan directamente con los datos subyacentes altamente dimensionales. Este es especialmente cierto para tareas bien estudiadas como reconocimiento de imágenes. Para tareas más complejas como el modelaje de estructuras sociales basadas en los CDR, el aprendizaje profundo se mantiene como un área de investigación activa.

En el actual análisis, dos series de características son generadas para cada polígono de red:

- **Orientado a la casa del cliente.** Esta primera serie está basada en los registros de clientes quienes viven dentro del polígono, determinado por su ubicación inferida del hogar. Esto es similar a tomar una encuesta de personas en su residencia usual. Las características son calculadas primero por cliente, luego agregadas al polígono de red. Ellas incluyen medidas de consumo (es decir, el número de llamada hechas o recibidas) y medidas de movilidad (es decir, dónde, qué lejos y con qué frecuencia viaja un cliente). Los datos son agregados al tomar el medio sobre todos los clientes, si es apropiado o sino una variedad de valores umbrales —por ejemplo, el número de clientes viviendo en un polígono de red quienes viajan regularmente más allá de 80 kilómetros de su hogar—.
- **Orientado a la activad de polígonos.** La segunda serie está basada en la actividad que ocurre dentro del polígono, sin importar dónde se ubica el hogar del cliente involucrado. Esto es similar a tomar una encuesta de personas que pasan por un área. En este caso, las características basadas en actividad incluyen el número de clientes que entran al polígono de la red, pero viven fuera del mismo, la frecuencia con la que lo visitan y el volumen de llamadas entrantes y salientes que son procesadas. Estas características son calculadas directamente al nivel de la red de polígonos y no requieren futuras agregaciones.

Ajuste del modelo

Estimar los índices de pobreza desde las características de CDR es un ejemplo de un problema de aprendizaje de máquinas "supervisado", o uno donde los datos entrantes son usados para predecir resultados conocidos —en este caso, las características de CDR e índices de pobreza—. Una vez que el modelo es construido, puede ser aplicado a nuevos datos de ingreso para los cuales el resultado correspondiente es desconocido. En este caso, el resultado desconocido sería geografías diferentes o diferentes puntos en el tiempo.

Los problemas de aprendizaje de máquinas supervisados son basados en clasificación, en dicho caso la variable resultante es una de una serie discreta de clases (es decir, masculino/femenino, pobre/no pobre, etc.), o basado en regresión, en la cual la variable resultante es un número continuo real expresado como un decimal, coeficiente o porcentaje. Los índices de pobreza son mayormente modelados naturalmente como una variable resultante continua. Sin embargo, también es

posible agrupar los datos en una serie pequeña de clases reflejando índices de pobreza bajos, moderados o altos. Ambos enfoques fueron examinados en este ejemplo, el cual probó una variedad de escenarios al mezclar diferentes series de datos de entrenamiento y prueba, y empleando tantos métodos de regresión y clasificación. La figura número 5 ilustra las diferentes combinaciones de datos y metodología[9].

Figura 5. Combinaciones de datos y metodología probados

Datos del teléfono celular		Encuesta de entrenamiento	Encuesta de prueba		Método		
2013	X	2006 total	2006 total	X	Regresión		
		2006 rural	2006 rural		Clasificación:		
		2006 urbano	2006 urbano		Ancho igual Igual probable Medios-K Basado en características	X	Línea base SVM Bosques aleatorios Aumento del gradiente Estocástico Medios-K Mezcla gaussiana Modelos de temas supervisados

Fuente: adaptado de Hong y Frías-Martínez (2015).

Caja 2. Aplicando los modelos de pobreza basados en los CDR para rellenar huecos de datos

Mucha de la discusión alrededor del uso de los CDR para la investigación socioeconómica ignora las condiciones prácticas en las cuales serán aplicadas. Los datos de los CDR están casi siempre disponibles adicionalmente a los datos convencionales, tales como encuestas de hogares o censos. De poco sirve mostrar que los datos CDR pueden ser usados para predecir estos datos de encuestas existentes; en vez de eso, el modelo ajustado debe ser aplicado a datos nuevos que no se ven para responder preguntas nuevas. Existen al menos tres diferentes maneras en las cuales los datos de CDR pueden complementar las fuentes convencionales de datos:

9. Para más información, ver: Hong y Frías-Martínez (2015).

1. **Relleno espacial: generando estadísticas de áreas pequeñas.** Dada la relativamente alta resolución espacial de CDR, el relleno espacial posiblemente ofrece el mayor valor agregado como un método de investigación socioeconómica. Una encuesta de hogar particular con un limitado tamaño de muestra solamente puede apoyar los estimados en una resolución espacial relativamente áspera, tal como al nivel del departamento. Las señales de comportamiento de alta resolución en los CDR pueden mejorar la fortaleza estadística de los datos de encuestas, permitiendo estimados precisos y más detallados. Idealmente, el periodo de recolección de CDR debe coincidir con el periodo en el que la encuesta fuese conducida. Este enfoque implica el riesgo de ignorar el rol que el espacio y la geografía puedan jugar en la predicción de índices de pobreza.
2. **Interpolación/extrapolación de tiempo.** La alta frecuencia potencial de los CDR puede también complementar las fuentes convencionales de datos. Las encuestas de hogares son actualizadas generalmente en intervalos de 2 a 5 años. Los CDR pueden ser usados para actualizar estos estimados, proveyendo a los oficiales con información actual sobre temas específicos de políticas. Un modelo predictivo puede ser construido para un año de encuesta usando datos CDR contemporáneos. Este modelo puede entonces ser aplicados a datos CDR más recientes para los cuales los datos de encuestas correspondientes no están disponibles. Existe un riesgo, sin embargo, de que la relación entre señales de comportamiento CDR y el objetivo variable, en este caso el índice de pobreza, pudiera cambiar con el tiempo.
3. **Extrapolación espacial.** Esta es la más ambiciosa aplicación potencial de los datos de CDR. En países o regiones en los cuales no hay datos recientes de encuestas disponibles, tal como áreas afectadas por conflictos o países que han experimentado inestabilidad política severa, los estimados pueden ser generados al usar datos de encuestas y CDR para una ubicación similar y luego aplicar este modelo a los datos CDR de la ubicación objetivo. Este enfoque requiere asunciones significativas, pero podría ser útil en casos donde no existe una fuente fuerte de datos. Hasta la fecha, la investigación sobre las aplicaciones prácticas de la extrapolación espacial ha sido limitada.

Evaluando el modelo

Una gran desventaja de los modelos de aprendizaje de máquinas es un fenómeno conocido como sobreajuste (*overfitting*). Los datos de entrenamiento siempre reflejan "señales" significativas y un "ruido" aleatorio. El sobreajuste ocurre cuando el modelo que se acopla tiene muchos parámetros libres por lo que se acopla tanto a la señal como al ruido. En dicho caso, el modelo parecerá ser un acople excelente para los datos de entrenamiento, con alto R^2, pero tendrá un rendimiento pobre cuando se aplica a datos externos a la muestra.

En el caso de Guatemala, una técnica llamada "validación cruzada" fue usada para protegerse del sobreajuste. La validación cruzada divide los datos de entrenamiento en partes, entrenando al modelo en una subserie de los datos (75%) y luego probándolo en los datos restantes (25%). Esto permite que los valores diagnosticados de la serie de prueba provean una muestra más precisa de cómo el modelo rendiría en un escenario fuera de la muestra. Las subseries de validación cruzada pueden ser construidas en múltiples ocasiones en diferentes maneras, con resultados promediados, para proveer mejores resultados.

Para evaluar los resultados de regresión, la precisión del modelo de aprendizaje máquina es medida usando R^2, raíz cuadrada del error cuadrático medio y la correlación entre los valores reales y pronosticados. R^2 mide la extensión en la cual el modelo explica la variabilidad de los datos de respuesta sobre su media, mientras que la raíz cuadrada del error cuadrático medio indica la diferencia entre los valores reales y los valores pronosticados. La calidad de las técnicas de clasificación es analizada usando dos medidas; precisión y valor F1. La precisión refleja el porcentaje de las muestras probadas cuya clase pronosticada es la misma que sus muestras reales. El valor F1 es un parámetro que combina la precisión y la cobertura del método, es decir, el número de muestras que están correctamente clasificadas y el número de muestras para las cuales se proporciona una etiqueta. En general, las metodologías más fuertes tienen valores mayores tanto para precisión como para la sensibilidad.

Resultados

Todos los modelos basados en CDR exhibieron un grado significativo de valor predictivo. Sin embargo, las especificaciones de diferentes modelos influenciaron en cómo de bien estas predecían índices de pobreza. Los análisis arrojaron cuatro resultados generales:

Resultado #1. Los CDR pueden predecir los índices de pobreza en Guatemala

A través de todos los modelos, el análisis CDR consistentemente predijo los índices de pobreza a nivel municipal, aunque su valor predictivo estaba limitado bajo ciertos parámetros experimentales. Los mejores modelos predijeron niveles de pobreza total en el 2006 con un R^2 de 0,76, indicando que aproximadamente 76% de la variación en índices de pobreza a nivel *municipal* podría ser explicada por los datos de teléfonos celulares de 2013, y con valoraciones F1 de hasta 0,84 para los modelos de clasificación, indicando que 84% de los municipios estaban clasificados según la categoría correcta cuando tres categorías separadas (índices de pobreza bajo, medio y alto) eran considerados. Por otra parte, los índices de predicción para los datos urbanos en 2006 mostraron valores R^2 de 0,69 para técnicas de regresión y 0,73 para clasificación, indicando un valor predictivo más débil. Los índices más bajos de predicción fueron para los datos rurales en el 2011, con resultados R^2 de 0,46 y 0,59 para modelos de regresión y clasificación, respectivamente. Los resultados de clasificación fueron ligeramente mejores que los resultados de regresión a través de todos los modelos debido al hecho de que clasificar índices de pobreza en tres clases es un problema de predicción más simple que tratar de aproximar valores reales. Experimentalmente, mientras más clases eran incluidas, las precisiones predictivas para la clasificación disminuyeron y convergieron con aquellas para la regresión. Por lo tanto, el número de clases de pobreza seleccionadas implican una compensación entre la precisión y la granularidad de la predicción.

Resultado #2. En Guatemala los CDR predicen la pobreza urbana y total con más precisión que la pobreza rural

Tanto en modelos de regresión como de clasificación, los índices de pobreza rural fueron consistentemente más difíciles de predecir que los índices totales o urbanos. Los valores R2 cayeron hasta casi 0,25 para pobreza rural, insinuando que los datos CDR podrían explicar solo el 25% de la variación en índices de pobreza rural. La precisión de los modelos de clasificación disminuyó hasta entre 0,3 y 0,65 dependiendo de la especificación. Dos hipótesis podrían explicar este fenómeno. Primero, los índices de penetración celular en áreas urbanas tienden a ser mayores, y por ende los análisis basados en CDR proveen señales de modelado más robustas en áreas urbanas, donde representan el comportamiento de una porción más grande de la población. En áreas rurales, menos teléfonos y menos llamadas debilitan la señal que puede ser extraída de los datos CDR, y su respectivo poder predictivo disminuye. Segundo, las áreas urbanas tienden a tener más antenas celulares por kilómetro cuadrado, lo cual resulta en polígonos

más pequeños. Los polígonos de mayor tamaño en áreas rurales pueden tender a reducir la granularidad de los datos al agregar comportamientos, debilitando el poder predictivo del algoritmo. Probar estás hipótesis requerirá de más investigación.

Resultado #3. Más análisis será necesario para determinar la extensión en la cual los modelos basados en CDR sobre datos de pobreza pasada pueden ser utilizados para predecir futuras dinámicas de pobreza

Los modelos basados en CDR podrían potencialmente ser usados para predicción temporal o extrapolación de tiempo (ver cajas 2 y 3), pero ninguna investigación ha podido todavía probar que esos modelos entrenados en valores de pobreza pasados pueden ser usados para predecir niveles de pobreza futuros. Dichos análisis requerirían una serie de datos extremadamente grandes y detallados. Por ejemplo, predecir los valores futuros de pobreza en Guatemala requeriría datos CDR del 2006 y 2011, además de los datos de encuestas de pobreza correspondientes para los mismos periodos de tiempo. Sin embargo, este análisis fue basado en datos de CDR de 2013 y niveles de pobreza rural para 2006 y 2011, y ningún dato de pobreza urbana o nacional fueron provistos. Como resultado, el modelo predictivo fue entrenado con los datos de CDR de 2013 y los datos de pobreza rural de 2006, mientras que los datos de CDR de 2013 fueron usados para predecir niveles de pobreza rural en el 2011. Mientras que este es el mejor enfoque metodológico dadas las restricciones de datos, los resultados preliminares mostraron valores bajos de R^2 de alrededor de 0,09 para regresión y valoraciones F1 máximas de 0,6.

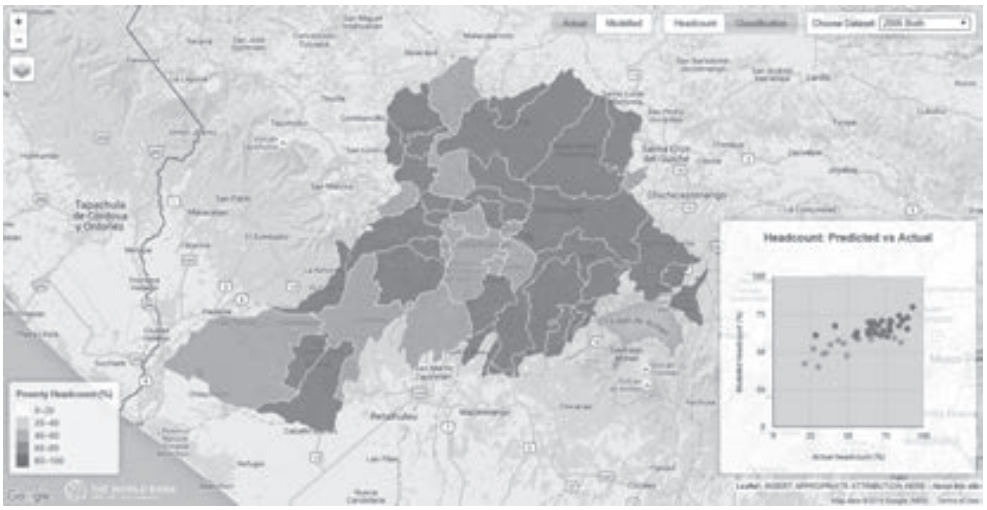
Caja 3. Usando visualización de datos interactivos para comunicar resultados de modelos

Los datos de Guatemala fueron mapeados en un sitio web interactivo. El mapa permitía a los usuarios cambiar entre modelos de regresión y clasificación, al igual que también diferentes periodos de datos. Los mapas muestran las estimaciones de pobreza mediante una escala de colores, mientras que los diagramas de dispersión y las matrices de confusión proporcionan evaluaciones más detalladas del rendimiento de cada modelo.

Figura 6. Índices actuales de pobreza (A) e índices de pobrezas modelados (B) para los municipios incluidos

(A)

(B)



Fuente: La visualización del autor basada en datos de Hong and Frías-Martínez (2015b).

Otras aplicaciones de datos de CDR para políticas de desarrollo

Mientras que los modelos de CDR no son aún lo suficientemente precisos para suplantar métodos tradicionales de investigación, las crecientes técnicas sofisticadas para combinar los CDR con otras fuentes de datos pudiera permitir que los datos celulares magnifiquen el valor de los censos y encuestas de hogares. Dos enfoques especialmente prometedores están descritos a continuación.

Predicciones a nivel de unidad usando encuestas móviles de datos reales del terreno

Un estudio reciente en Ruanda usó encuestas focalizadas de teléfono para recolectar datos sobre la riqueza personal en vez de la incidencia de pobreza[10]. La ventaja de este enfoque es que las respuestas de encuestas de teléfono pueden ser combinadas con los datos de CDR a nivel individual. Esto típicamente no es posible con encuestas de hogar oficiales, las cuales son usualmente anónimas y no contienen un número celular que pueda ser analizado con referencias cruzadas de los registros de CDR.

Al coordinar la recolección de respuestas de encuestas con análisis de CDR, el estudio pudo desarrollar una serie de datos individuales de gran calidad. Sin embargo, el uso de encuestas de teléfono para recolectar datos presenta sus propias limitaciones, ya que no es posible obtener la misma información de consumo detallado que las encuestas de hogar cara a cara generalmente revelan. En vez de eso, los valores aproximados de riqueza deben ser usados —en este caso, con preguntas sobre propiedad de activos—. Si los resultados no son cuidadosamente validados, esto puede llevar a errores en las variables resultantes. Adicionalmente, emparejar los datos con las encuestas de teléfonos requiere que los CDR no sean anónimos, lo cual genera preocupaciones sustanciales sobre privacidad.

El estudio en Ruanda ofreció una prueba motivadora de concepto para la aplicación de datos de CDR para la "extrapolación de tiempo". Encontró que un modelo de CDR para 2009 proveyó una evaluación de la pobreza más precisa que la encuesta de hogar obsoleta de 2007. Sin embargo, el modelo CDR no era únicamente apropiado para esta tarea, ya que los indicadores macroeconómicos como el crecimiento del PIB podría ser usado para crear una proyección similar precisa[11]. Además, el modelo CDR solo reproducía parcialmente los estimados de la riqueza de hogares promedio a nivel de distrito registrados por las encuestas de hogar. Aunque correlaciones de cerca de 0,9 fueron reportadas a nivel de distrito —mayor que en el análisis de Guatemala— parece que unos pocos distritos adinerados pueden haber generado los resultados obtenidos.

10. Blumenstock *et al*. (2015).
11. Ver Beegle (2016).

Combinando los datos de CDR con datos observables por el público

Durante los últimos años el proyecto WorldPop[12], con base en la Universidad de Southampton, ha producido mapas cuadriculados de la población para decenas de países, los cuales detallan la población estimada por cada 100 metros cuadrados[13]. WorldPop también ha producido estimaciones de pobreza cuadriculadas a una resolución de 1 kilómetro para un pequeño número de países usando un método similar basado en datos de encuestas de hogares. En colaboración con la Fundación Flowminder, WorldPop está ahora empezando a integrar características derivadas de CDR dentro del mismo marco. Mientras que relativamente poca información ha sido publicada sobre este enfoque, tiene la ventaja de poder incorporar todos los datos relevantes en un mismo formato. Los mapas resultantes deben no solo ser altamente precisos, sino también tan consistentes como sea posible con las series de datos subyacentes. Una advertencia sobre el enfoque WorldPop es que la inclusión de la dimensión de pobreza aún está en una etapa inicial, y no está claro cómo estos mapas son validados, ya que por diseño deberían correlacionar con las estimaciones de pobreza basadas en encuestas[14].

› Usar datos de CDR para trazar el nivel de la pobreza

Mientras que los enfoques técnicos del uso de los CDR para trazar el nivel de la pobreza continúan evolucionando, se requiere abordar un número de retos para poder poner en práctica esta metodología como una herramienta práctica para el análisis de políticas. Varios de estos retos están descritos abajo. Un rango de otros problemas éticos y legales está considerado en un informe oficial (Libro Blanco) realizado entre el Banco Mundial y Data-Pop Alliance[15].

Validación y transparencia

La validación es más complicada para modelos de CDR de lo que es para técnicas basadas en encuestas. Los instrumentos de encuestas son relativamente

12. Http://www.worldpop.org.uk
13. Estos mapas son construidos al combinar datos de censos con covariables físicas, como clima, elevación, inclinación y cuerpos de agua, y covariables humanas, tales como expansión humana, basada en imágenes de satélite, asentamientos conocidos, caminos y puntos de interés bajo un marco Bayesiano.
14. El Laboratorio de Innovación del Banco Mundial está patrocinando actualmente trabajos de análisis CDR adicionales con Flowminder/Worldpop en Haití, lo cual se espera que proveerá mayor introspectiva sobre sus métodos.
15. Letouzé y Vinck (2015).

transparentes. Estas pueden ser inspeccionadas, y el trabajo de campo cuantitativo y cualitativo realizado antes y después de la recolección de datos puede generar confianza. Las inconsistencias temporales y espaciales pueden ser revisadas y en raros casos las encuestas superpuestas pueden ser verificadas. Como resultados, los programas de encuestas como el Estudio de Medición de Estándares de Vida y Encuestas de Salud son usualmente tratados como lo más cercano a la realidad del terreno.

La validación es generalmente más difícil para investigaciones de grandes datos y para modelos de CDR en particular. Estos datos no son recolectados específicamente para análisis socioeconómicos, por lo que la inspección puede no ser útil y pruebas previas no son usualmente posibles. Además, los modelos de CDR son diseñados usualmente para interpolar entre modelos convencionales de encuesta bien sea en tiempo o espacio, por lo que la validación directa contra una encuesta no es posible.

Sin embargo, las estrategias de validación rigurosas deben ser desarrolladas. Como mínimo, las técnicas de validación dentro de muestras deben ser usadas[16]. El potencial para la validación fuera de muestras no se conoce, pero el acceso a las series de datos de CDR más largos, los cuales incluyen más de una encuesta de hogar, pudiera ayudar a lidiar con esta preocupación.

Privacidad

La privacidad de datos es un problema sensible y usualmente controvertido, por lo que ciertas precauciones deben ser tomadas antes de analizar datos de CDR. Los identificadores deben ser oscurecidos antes que los registros sean exportados desde los sistemas de las MNO, para que los números de teléfonos celulares o campos similares no permanezcan en los datos salientes. Este proceso es referido como "seudonimización". En general, la seudonimización debe asegurar que el mismo seudónimo sea aplicado a la historia completa de llamadas de un usuario dado.

La investigación sugiere que la seudonimización simple puede ser revertida por un determinado individuo armado con información auxiliar relativamente fácil de obtener, tal como las direcciones de trabajo y hogar de una persona y una o dos ubicaciones a las que se conoce que visitan en momentos particulares[17]. Por

16. Las técnicas de validación dentro de muestras incluyen series de pruebas de retención o validación en cruce de descartar-uno.
17. De Montjoye (2013).

lo tanto, las medidas técnicas e institucionales adicionales son necesarias para restringir el acceso a los datos. Esto generalmente implica acuerdos de no divulgación entre las MNO y los investigadores, y las investigaciones, desarrollos y comunidades MNO se están esforzando para racionalizar este proceso.

Adicionalmente, cualquier dato final como resultado del análisis CDR no debe comprometer la privacidad del individuo. Las precauciones similares a aquellas usadas cuando se liberan tabulaciones de los datos de encuestas tradicionales puede reforzar la privacidad de los datos de CDR. Estas medidas pueden incluir agrupar datos donde los tamaños de las células caen debajo de un número fijo de personas (es decir, 5 o 10), o variables de codificación altos y bajos en casos donde los valores extremos pudieran ser reveladores.

Otras aplicaciones prometedoras de la analítica CDR

Los CDR ofrecen una rica serie de datos para estudiar poblaciones, y aplicaciones numerosas están emergiendo en áreas más allá de la medición de la pobreza. Dos aplicaciones de relevancia particular al trabajo del Banco Mundial en Guatemala y la región de Latinoamérica están descritas a continuación[18].

Análisis de transporte

El análisis de transporte está entre las aplicaciones más prometedoras para los modelos de CDR. Las características económicas y sociales son solo registradas implícitamente en los datos de CDR, y por ende fuertes asunciones y modelajes complejos son requeridos para inferirlos. El comportamiento espacial, por otra parte, está explícitamente registrado por las ubicaciones de las antenas celulares. Por esta razón el análisis puramente espacial tiende a ser más simple y más robusto que otras formas de análisis de CDR[19].

Debido a la relativa simplicidad del análisis, algunas MNO y terceras partes están empezando a ofrecer productos de datos de transporte estándar a gobiernos locales y nacionales.

18. Una reseña más completa es provista en Blondel (2015).
19. Por ejemplo, Angelakis *et al.* (2013) usó datos CDR para examinar los patrones de transporte Cote d'Ivoire. Ellos descubrieron que varias matrices de viajes, incluyendo rutas y horas de viajes podrían ser calculados desde esos datos tanto a nivel nacional y a nivel de ciudad.

Preparación y respuesta ante desastres

Los grandes movimientos de la población usualmente ocurren a raíz de los desastres naturales, y estos movimientos pueden dejar tanto a los censos como las encuestas en lugares de predesastre en forma obsoleta. Para poder proveer asistencia humanitaria y restaurar servicios básicos en las áreas afectadas por el desastre, los gobiernos y las organizaciones requieren de datos actualizados de la población que puedan ser recolectados rápidamente y a un costo modesto. Este fue el caso de Haití después del terremoto de 2010, el cual inspiró a varios investigadores a examinar el potencial de datos de CDR para producir rápidamente información de rastreo de alta frecuencia sobre el desplazamiento de la población a corto plazo. Además, se encontró que esos datos históricos detallados sobre la movilidad de la población en áreas de predesastre podrían ser usados para predecir respuestas de residentes al terremoto, permitiéndole a las agencias prepararse mejor para futuros desastres[20]. Las técnicas similares han sido desde entonces aplicadas exitosamente durante y después de otros desastres naturales[21].

Conclusiones

La expansión dramática del uso de teléfonos móviles en países en desarrollo en años recientes ha producido una fuente de información rica y mayormente sin explotar sobre las características de las comunidades y regiones. Los métodos de investigación basados en CDR tienen el potencial para proveer estimados detallados y confiables de índices de pobreza en tiempo real y a un costo mucho menor que las encuestas tradicionales. Estos métodos tienen aplicaciones especialmente prometedoras en países en desarrollo, como en Guatemala, donde los altos índices de pobreza e inequidad y los limitados recursos fiscales y presupuestarios complican la tarea de recolectar datos y acentuar la importancia de focalizar con precisión el gasto público.

El análisis de CDR puede complementar métodos convencionales de investigación al mejorar la fortaleza estadística de datos de encuestas y al extrapolar estos datos a través del espacio y tiempo. El análisis presentado anteriormente estaba limitado a datos de CDR agregados y cifrados de solo cinco departamentos administrativos en el suroeste de Guatemala, y los resultados sugieren que expandir el tamaño de la muestra permitiría estimados de pobrezas más

20. Flowminder (2016a).
21. Ver para ejemplo Moumny *et al.* (2013) y Flowminder (2016b).

robustos y confiables. Los legisladores en Guatemala podrían obtener series de datos más exhaustivos al trabajar directamente con las MNO.

Mientras las metodologías analíticas son desarrolladas, las aplicaciones CDR podrían extenderse más allá del estudio de la pobreza. Los CDR podrían permitir a los legisladores rastrear patrones de crimen, inseguridad de comida, enfermedades epidémicas y otros problemas sociales y económicos en tiempo real. Los legisladores en Guatemala y otros países en desarrollo ahora tienen la capacidad de acceder a una fuente de riqueza de datos celulares. El establecer fuertes asociaciones con operadoras móviles será el primer paso para aprovechar el enorme potencial de los datos celulares para la investigación socioeconómica y análisis de la política.

Referencias bibliográficas

Angelakis, V., Gundlegård, D., Rajna, B., Rydergren, C., Vrotsou, K., Carlsson, R., Forgeat, J., Hu, TH., Liu EL., Moritz, S., Zhao, S., Zheng, Y. (2013). Mobility modeling for transport efficiency- analysis of travel characteristics based on mobile phone data. In: *Mobile phone data for development-analysis of mobile phone datasets for the development of Ivory Coast*. Orange D4D challenge, pp. 412-422.

Beegle, K., Christiaensen, L., Dabalen, A., Gaddis, I. (2016). *Poverty in a Rising Africa*. Washington, DC: World Bank. doi: 10.1596/978-1-4648-0723-7.

Blondel, Vicent D., Decuyper, A., Krings, G. (2015). "A survey of results on mobile phone datasets analysis". *EJP Data Science*. http: //link.springer.com/ article/10.1140/ epjds/s13688-015-0046-0.

Blumenstock, J., Cadamuro, G., On, R. (2015). "Predicting poverty and wealth from mobile phone metadata", *Science,* 350: 1073-1076.

Elbers, Ch., Lanjouw, P., Lanjouw, J. (2003). "Micro-Level Estimation of Poverty and Inequality". *Econometrica*, vol. 71 (1): 355-364.

Flowminder (2016a). "Case Study: Haiti Earthquake 2010", http: //www. flowminder. org/case-studies/haiti-earthquake-2010

Flowminder (2016b). "Case Study: Nepal Earthquake 2015", http: //www. flowminder. org/case-studies/nepal-earthquake-2015

Frías-Martínez, V., Frías-Martínez, E., Oliver, N. (2010). "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records". https: // www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewFile/1094/1347

Frías-Martínez, V., Virseda, J. (2012). "On the relationship between economic factors and cell phone usage", International Conference on Technologies and Development, ICTD.

Hong, L., Frías-Martínez, E., Frías-Martínez, V. (2016). "Topic Models to Infer Socio-economic levels", Thirtieth International Conference on Artificial Intelligence, AAAI.

Hong, L., Frías-Martínez, V. (2015a). "Estimating Incidence Values Using Mobile Phone Data: Deliverable 2: Statistical Models". Unpublished manuscript, June 4.

Hong, L., Frías-Martínez, V. (2015b). "Prediction of Incidence Levels". Unpublished manuscript, June 4.

Letouzé, E., Vinck, P. (2015). "The Law, Politics and Ethics of Cell Phone Data Analytics". Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, April.

Montjoye, Y.-A. De, Hidalgo, C. A., Verleysen, M., Blondel, V. D. (2013). "Unique in the Crowd: The privacy bounds of human mobility". *Nature S.Rep.* 3.

Moumny, Y., Frías-Martínez, V., Frías-Martínez, E. (2013). "Characterizing Social Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake", Third Workshop on Pervasive Urban Applications @ Pervasive'13, Zurich, Switzerland.

Sánchez, S., Scott, K., López, H. (2016). *Guatemala: Closing Gaps to Generate More Inclusive Growth*. Washington, D.C.: The World Bank.

Simon, P. (2012). "IFC Mobile Money Scoping: Country Report: Guatemala", International Financial Corporation, World Bank Group, http: //www.ifc.org/wps/wcm/ connect/8b233f0043efb60d95b6bd869243d457/Guatemala+Public.pdf?MOD=AJPERES