

# Segmentación de la población internauta en Bolivia

## 1. Introducción:

Nos encontramos en una nueva etapa de desarrollo tecnológico donde la tecnología ya deja de ser un problema, mientras el saber qué hacer con ella y los datos que la misma genera se convierte en tema central de debate.

Hoy en día es muy normal que las empresas empleen técnicas de segmentación de mercado para conseguir clientes o incrementar las ganancias. Para realizar un análisis de segmentación es preciso tener claro qué tipo de salidas se quiere obtener dependiendo de la necesidad del “negocio”. La segmentación de mercado divide un mercado en segmentos más pequeños de compradores que tienen diferentes necesidades, características y comportamientos que requieren estrategias o mezclas de marketing diferenciadas. Esta técnica no solo es aplicada en temas de marketing empresarial, tal es el caso de la empresa Cambridge Analítica una empresa de datos digitales que cuenta con información personal de más de 230 millones de votantes en Estados Unidos que trabajó para la campaña electoral del candidato republicano Donald Trump dividiendo a las audiencias publicitarias en grupos pequeños, para que, posteriormente, se les dirijan anuncios a través de “múltiples plataformas”. Si bien los datos pueden ayudar en campañas publicitarias o incremento de ventas en caso de empresas con fines de lucro, en el ámbito gubernamental podría coadyuvar en la creación de políticas de estado para apoyar a sectores vulnerables según las necesidades que presenten en temas relacionados a servicios básicos, acceso a la educación, acceso a la tecnología e internet.

El presente trabajo pretende encontrar un prototipo de modelo que permita segmentar los registros de la encuesta de Tics realizada por la Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación.

### ***Palabras clave:***

*Segmentación<sup>1</sup>, internauta<sup>2</sup>, dataset<sup>3</sup>, variables características<sup>4</sup>, cluster<sup>5</sup>.*

---

<sup>1</sup> **Segmentación**, división una población en conjuntos más pequeños que tienen diferentes necesidades, características y comportamientos.

## 2. Objetivo

Identificar segmentos de grupos sociales en la población boliviana internauta de tal forma que se elaboren políticas de estado en temas relacionados a tecnología, uso de redes sociales, educación, telecomunicaciones y servicios básicos para fortalecer la soberanía tecnológica la cual es pilar fundamental de la agenda 2025.

### 2.1. Objetivos específicos:

- Identificar la población y las fuentes de información
- Identificar las variables características para la conformación del dataset
- Aplicar procesos de limpieza y transformación al dataset
- Encontrar el modelo que permita identificar los segmentos en la población
- Analizar los segmentos encontrados y emitir conclusiones/recomendaciones

## 3. Limitaciones:

No se consideran todas las variables por su varianza o cantidad de datos faltantes. Los segmentos encontrados son un aproximado de la realidad toda vez que depende del nivel de confiabilidad de la encuesta y del modelo encontrado. El modelo encontrado puede ser mejorado si se lo sigue alimentando.

---

<sup>2</sup> *Internauta*, persona que utiliza los servicios de internet u otra red informática

<sup>3</sup> *Dataset*, nominativo para un conjunto de datos tabulados.

<sup>4</sup> *Variables característica*, atributos de un registro o entidad, cuantitativo o cualitativo.

<sup>5</sup> *Cluster*, nominativo para segmento en procesos de análisis de datos.

## 4. Desarrollo

### 4.1. Identificar la población

Se emplea el dataset publicado por la Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación AGETIC en el portal de datos abiertos datos.gob.bo



Figura 1: Portal de datos abiertos de Bolivia

### Datos y Recursos






	<b>Base de Datos Encuesta TIC, 2016</b>	<a href="#">Explorar</a>
	<b>Base de Datos Encuesta TIC, 2016</b>	<a href="#">Explorar</a>
	<b>Diccionario de Datos</b>	<a href="#">Explorar</a>
	<b>Cuestionario Internautas</b> Formulario de Encuesta	<a href="#">Explorar</a>
	<b>Cuestionario de NO Internautas</b> Formulario de Encuesta	<a href="#">Explorar</a>

Figura 2: Archivos de la encuesta Agetic

- Se emplea el archivo base-5536-bdfinalcorregido.csv como data set de entrada.
- Se emplea el archivo Diccionario de datos PDF para la comprensión de las columnas del dataset.
- Se emplea el archivo Cuestionario Internautas para comprender las preguntas de la encuesta y su relación con cada columna dentro del archivo base-5536-bdfinalcorregido.csv.

#### **4.2. Identificar las variables características para la conformación del dataset**

Se filtraron las preguntas del cuestionario a formar parte del modelo bajo los siguientes criterios:

- Variables que generalicen a la población, es decir, aquellas preguntas que en su mayoría hayan sido contestadas por los encuestados, por ejemplo, la pregunta 81. ¿Generalmente qué forma de pago usa para comprar en Internet?, no se considera debido a que menos de la mitad de la población realizó alguna compra en internet, y en su mayoría esta columna se encontrará con datos vacíos.
- Variables que puedan ser agrupadas en estratos y no tengan un nivel alto de varianza por ejemplo, la pregunta: 49. ¿Cuáles son las tres páginas web internacionales que más visita? Podría contener una varianza alta por temas de escritura, de ortografía y otros.

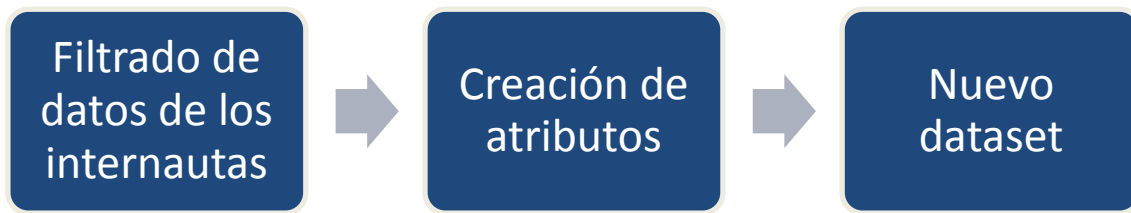
Según los puntos mencionados con anterioridad, se consideraron las siguientes preguntas del cuestionario para la conformación del dataset:

- Preguntas filtro 1, 2, 3
- Acceso y uso de equipamiento y servicios TIC 5, 6, 10, 11, 12, 15, 16, 17, 20, 23, 27, 28, 31, 32, 33, 36, 37, 41
- Hábitos en el acceso a internet 43, 46, 47, 69, 71
- Uso de internet para la actividad económica o laboral 73, 76
- Comercio electrónico 77, 86
- Banca por internet 95, 96, 97
- Redes sociales 99, 100, 101, 102, 105, 109, 110, 113, 117, 122, 125
- Habilidades informáticas 143, 144
- Socio demográficas y estratificación 148, 149, 150, 152, 153, 154, 155, 156, 157

### 4.3. Aplicar procesos de limpieza y transformación al dataset

Se empleó la herramienta de trabajo RapidMiner <sup>6</sup>(AGPLv3) para los procesos de limpieza y transformación de los datos.

El archivo generación-atributos-dataset.rmp contiene el archivo para RapidMiner con el que se realizó la transformación de atributos el cual sigue el siguiente flujo:

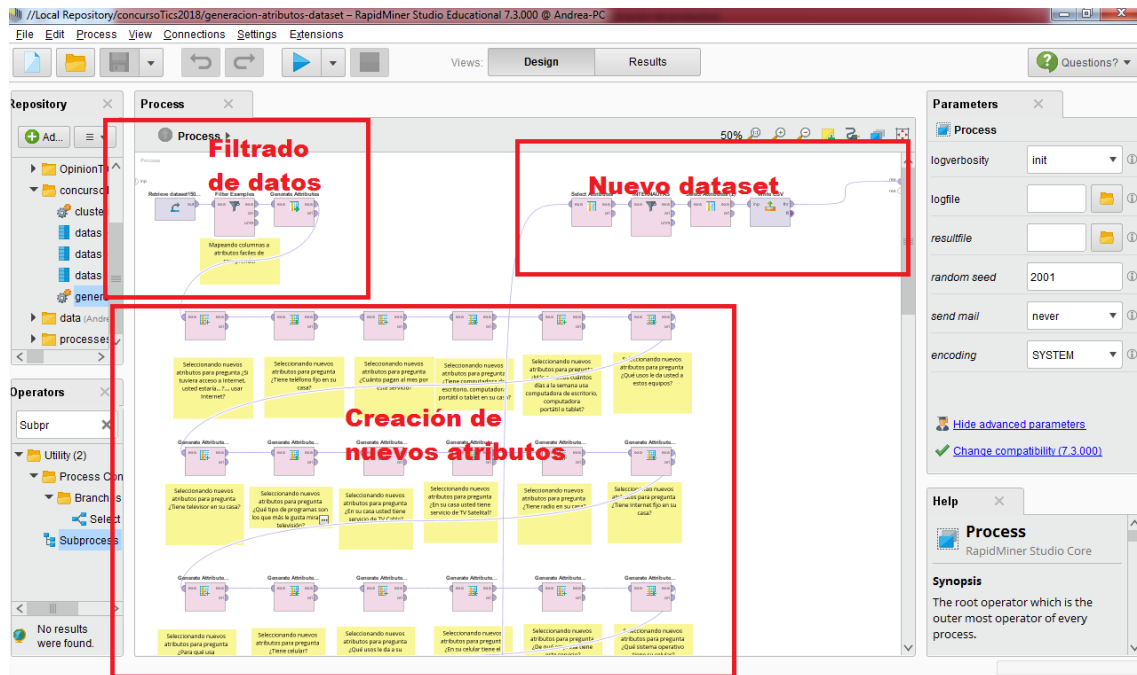


**Figura 3:** Proceso de limpieza y transformación

- Filtrado de datos de los internautas – Se leen los registros del dataset base-5536-bdfinalcorregido.csv y se filtran aquellos cuyo valor del P2 (2. ¿Usted ha navegado por Internet o se ha conectado a alguna red social a través de Internet?) es igual a 1 (SI)
- Creación de atributos – Se crean nuevos atributos binomiales (SI/NO, 1/0) con nombres representativos a partir de las columnas de la siguiente forma (Ejemplo):
  - P1 = edad
  - P4 = tiene\_telefono\_fijo
  - P101A/B/C/D/E = usa\_red\_social\_compartir\_contenidos, usa\_red\_social\_ver\_videos , etc.
- Nuevo dataset – El nuevo conjunto de atributos se almacena en el archivo dataset22042018.csv.

---

<sup>6</sup> **RapidMiner**, programa informático para el análisis y minería de datos.



**Figura 4:** Proceso de limpieza y transformación en RapidMiner

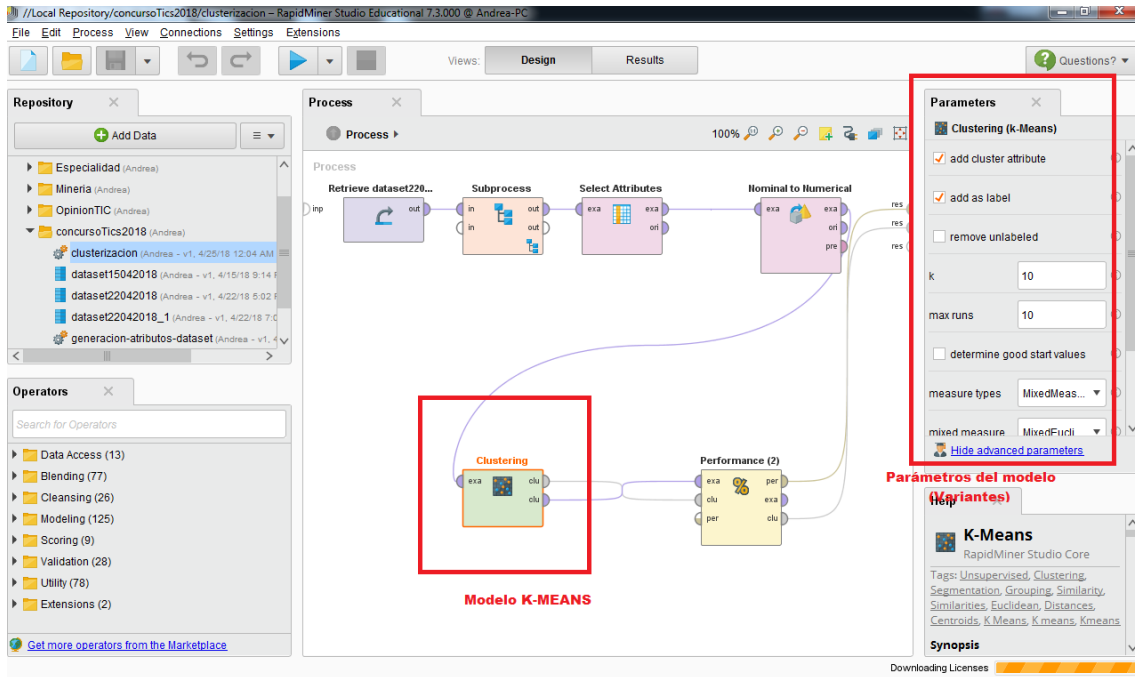
#### 4.4. Encontrar el modelo que permita identificar los segmentos en la población

Dado que el presente análisis sobre la encuesta de TICS cuenta con diferentes observaciones de la realidad y no se quiere predecir, sino agrupar o segmentar, se empleó aprendizaje no supervisado para su desarrollo.

Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones y no depende de un atributo a predecir o clasificar (Hinton & Sejnowski, 1999).

Se empleó el método de segmentación K-MEANS que es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de  $n$  observaciones en  $k$  grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano (Dhillon, 2001).

Se empleó la herramienta de trabajo RapidMiner para el proceso de entrenamiento del modelo K-MEANS. El archivo clusterizacion.rmp contiene los procesos para esta etapa.



**Figura 5:** Proceso entrenamiento del modelo en RapidMiner

En la figura anterior, se puede ver los componentes del proceso de entrenamiento, en el panel derecho se encuentran los parámetros empleados para el modelo donde K es el número de clusters que se quiso encontrar, este valor es variante y se lo elige según la mejor medida de efectividad que se obtiene en casa experimento.

La medida de efectividad empleada para encontrar el mejor modelo es la de Davies Boulding.

Davies Boulding es una medida para evaluar algoritmos de clusterizacion empleando la siguiente fórmula:

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^k \max \left( \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

**Figura 6:** Fórmula para cálculo de índice Davies Boulding

Donde k es el número de clústeres,  $\sigma_i$  es la distancia promedio entre cada punto en el clúster i y el centroide del clúster,  $\sigma_j$  es la distancia promedio entre cada punto del

clúster  $j$  y el centroide del clúster, y  $d(c_i, c_j)$  es la distancia entre los centroides de los 2 clústeres .

Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros están bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo.

Se realizaron diferentes experimentos obteniendo los siguientes resultados:

Cantidad de clusters/segmentos	Índice Davies Boulding
<b>10</b>	0.024
<b>30</b>	0.023
<b>40</b>	0.022
<b>50</b>	0.021
100	<b>0.018</b>
<b>200</b>	Infinito

**Tabla 1:** Comparación de modelos

Como se puede ver en la tabla anterior, el mejor valor se obtiene cuando se aplican 100 clusters (IDB= 0.018) ya que con 200, el IDB se dispara (Converge).

#### 4.5. Analizar los segmentos encontrados

Una vez identificados los clusters, se analiza el centroide del grupo, el cual es un registro representativo dentro del mismo, se asume que las características de este registro representa al segmento o perfil que representa, las cuales lo diferencian de otros perfiles/segmentos.

En el archivo adjunto segmentación.xlsx se encuentran las características de los centroides de los segmentos identificados como cabeceras de las filas y las variables características como columnas. El valor de la celda (Entre 0 y 1) representa la probabilidad del registro.



Atributo	cluster0	cluster1	cluster2
alcantarillado = 0	0.2890295358649789	0.03623188	0.04347826
alcantarillado = 1	0.7109704641350211	0.96376811	0.95652173
apoyo_causa_solidaria = 0	0.7974683544303798	0.89613526	0.86956521
apoyo_causa_solidaria = 1	0.9050632911392406	0.86231884	0.78743961
busca_ciencia_tecnologia_internet	0.37763713080168776	0.22946859	0.17149758
busca_educacion_internet	0.4430379746835443	0.44927536	0.24637681
busca_entretenimiento_farandula_internet	0.6118143459915611	0.43719806	0.23429951
busca_noticias_en_internet	0.5147679324894515	0.30193236	0.56038647
calefon_o_ducha = 0	0.7426160337552743	0.44685990	0.22946859
calefon_o_ducha = 1	0.25738396624472576	0.55314009	0.77053140
casa_propia = 0	0.2848101265822785	0.25845410	0.30676328
casa_propia = 1	0.7151898724177716	0.74154589	0.69323671

**Figura 7:** Atributos de los centroides

La forma de interpretación se realiza según la probabilidad del atributo, por ejemplo:

Atributo	cluster0
usa_facebook	0.9852320675105485

**Figura 8:** Ejemplo de interpretación probabilidad alta

Los miembros del cluster0 presentan una probabilidad del 0.98 (Muy Alta) de usar Facebook.

Atributo	cluster0
tarjeta_credito = 1	0.029535864978902954

**Figura 9:** Ejemplo de interpretación probabilidad baja

Los miembros del cluster0 presentan una probabilidad de 0.029 (Muy baja) de tener tarjeta de crédito (Donde = 1 representa SI y = 0 representa NO).

Si bien se encontró que el mejor modelo es la segmentación con 100 grupos, por fines demostrativos a continuación se describirán los clusters que mayor relevancia sugieren con una segmentación de 40 cluster en el archivo segmentacion-filtro.xlsx.

*Nota.- Se deja a disposición de otros analistas los demás clusters encontrados (10, 30, 40, 50 y 100) en el archivo segmentacion.xlsx*

Bajo esta estrategia, se han identificado los siguientes grupos relevantes:

<b>CLUSTER 5</b>	
<b>Características</b>	<ul style="list-style-type: none"><li>- Tendencia a apoyar causas políticas por redes sociales</li><li>- Cliente leal de Tigo</li><li>- Alto uso de Youtube</li><li>- La mayoría son jóvenes</li><li>- No genera ingresos</li><li>- No tiene pareja</li><li>- Llego a la universidad y colegio</li><li>- Jefe hogar llego a ser profesional</li><li>- Cuenta con los servicios básicos</li><li>- Tiende a buscar información en internet</li><li>- Ingresos económicos en hogar regulares</li></ul>
<b>Comentario</b>	Se resalta a Tigo como empresa preferida, la tendencia a apoyar causas políticas en redes sociales y en su mayoría jóvenes.

**Tabla 2: Segmento 5**

<b>CLUSTER 7</b>	
<b>Características</b>	<ul style="list-style-type: none"><li>- Consulta movimientos bancarios, transferencias o pago de servicios por cuenta bancaria</li><li>- Cuenta con tarjeta de crédito</li><li>- Tendencia a realizar viajes al exterior</li><li>- Tendencia a tener un plan post pago</li><li>- Tendencia a hacer negocios en redes sociales</li><li>- En su mayoría adultos</li><li>- Tiende a ingresos altos</li><li>- Cuenta con los servicios básicos</li><li>- Llego a ser profesional</li><li>- Jefe del hogar llego a ser profesional</li><li>- Tiende a familia reducida</li></ul>
<b>Comentario</b>	Se resalta la alta interacción con el banco, la tendencia a viajar, a buscar un plan Post Pago, la tendencia a percibir ingresos altos y la profesión al igual que el jefe de hogar.

**Tabla 3: Segmento 7**

---

### CLUSTER 14

<b>Características</b>	<ul style="list-style-type: none"><li>- Trabajo relacionado a tecnología</li><li>- Consulta movimientos bancarios, transferencias, pago de servicios por cuenta bancaria</li><li>- Conoció personas por internet</li><li>- Tendencia a usa twitter</li><li>- Tendencia a hacer negocios en redes sociales</li><li>- Mayoría son varones</li><li>- Genera ingresos</li><li>- Llegó a la universidad</li><li>- Cuenta con los servicios básicos</li><li>- Jefe de hogar llegó a la universidad</li><li>- Tiende a familia reducida</li><li>- Usa internet la mayor parte de la semana</li></ul>
<b>Comentario</b>	Se resalta el trabajo con tecnología, la interacción con banco, uso de twitter y que sean en su mayoría varones.

---

**Tabla 4:** Segmento 14

---

### CLUSTER 20

<b>Características</b>	<ul style="list-style-type: none"><li>- Está tendiendo a no usar SMS por celular</li><li>- Mas mujeres</li><li>- Joven</li><li>- Ingresos del hogar bajos</li><li>- No tiene pareja</li><li>- Llegó a la Universidad</li><li>- Jefe de hogar llegó a colegio</li><li>- Bajo uso de celular o internet para buscar información</li></ul>
<b>Comentario</b>	Se resalta la tendencia a ingresos económicos bajos y en su mayoría son mujeres.

---

**Tabla 5:** Segmento 20

---

### CLUSTER 21

<b>Características</b>	<ul style="list-style-type: none"><li>- Tiende a no usar su celular para internet</li><li>- Tendencia a plan post pago</li><li>- Tendencia a no usar ANDROID como Sistema operativo</li><li>- Tiende a no usar Facebook ni Whatsapp</li><li>- Desconfía de las redes sociales</li><li>- No tiene pareja</li><li>- Llegó a colegio o universidad</li><li>- Jefe de hogar llegó a colegio</li><li>- Ingresos económicos en el hogar de bajos a regulares</li><li>- Joven o adulto</li></ul>
------------------------	---

<b>Comentario</b>	Se resalta la tendencia a alejarse del internet, redes sociales y tecnología.
-------------------	---

---

**Tabla 6:** Segmento 21

---

### CLUSTER 22

<b>Características</b>	<ul style="list-style-type: none"><li>- Tendencia alta a tener internet fijo</li><li>- Tendencia alta a buscar información en internet</li><li>- Tendencia a usar internet con fines académicos</li><li>- Tendencia a usar redes sociales con fines académicos</li><li>- Alto uso de Youtube</li><li>- Tendencia a usar twitter</li><li>- Joven o colegial</li><li>- Tiende a ingresos altos en el hogar</li><li>- No genera ingreso</li><li>- No tiene pareja</li><li>- Llego a colegio o universidad</li></ul>
------------------------	--

<b>Comentario</b>	Se resalta la tendencia a buscar información en internet y redes sociales con fines académicos y la tendencia a percibir ingresos económicos altos.
-------------------	---

---

**Tabla 7:** Segmento 22

---

## CLUSTER 32

<b>Características</b>	<ul style="list-style-type: none"><li>- Tendencia a usar VIVA</li><li>- En su mayoría varones</li><li>- Joven</li><li>- No Genera ingresos</li><li>- No tiene pareja</li><li>- Llegó a colegio o universidad</li><li>- Jefe de hogar llegó a la universidad</li><li>- Tiende a familia amplia</li><li>- Tiende a buscar entretenimiento en internet</li><li>- Ingresos económicos en hogar regulares</li></ul>
<b>Comentario</b>	Se resalta VIVA como empresa preferida en Telecomunicaciones, en su mayoría jóvenes varones.

---

**Tabla 8:** Segmento 32

### Conclusiones

Se logró identificar 8 segmentos significativos y diversos en características, se resalta atributos como el género, ingresos económicos, nivel de instrucción, intereses, interacción con la tecnología y preferencias en empresas de telecomunicación.

### Recomendaciones

Se recomienda ampliar trabajos como este en el ámbito de datos abiertos en Bolivia para proponer políticas de estado que coadyuven en temas sociales como:

- Impulsar el uso de las tecnologías e internet con fines académicos de forma que a futuro la persona pueda percibir ingresos económicos altos.
- Impulsar la inserción de la mujer en trabajos relacionados a tecnología.
- Impulsar la formación profesional aun cuando el jefe de hogar lo logró culminar el colegio en algunos casos.
- Impulsar el uso de las tecnologías e internet en aquellos sectores que desconfían de las mismas o no las conocen.

Así mismo se recomienda agregar variables demográficas, geográficas, psicográficas y de estilo de vida para tener una mejor segmentación y obtener un mejor conocimiento de la realidad de la sociedad internauta en Bolivia.