

**CLUSTERIZACIÓN DE ZONAS NO INTERCONECTADAS DE  
COLOMBIA ORIENTADA AL USO DE LA GASIFICACION DE  
BIOMASA COMO FUENTE ENERGETICA.**

NICOLAS ALBERTO MOJICA CRUZ  
YILBER MANUEL MÉNDEZ CALDERÓN

UNIVERSIDAD AUTONOMA DE BUCARAMANGA  
FACULTAD DE INGENIERIA  
PROGRAMA DE INGENIERIA EN ENERGÍA  
BUCARAMANGA

2021

**CLUSTERIZACIÓN DE ZONAS NO INTERCONECTADAS DE  
COLOMBIA ORIENTADA AL USO DE LA GASIFICACION DE  
BIOMASA COMO FUENTE ENERGETICA.**

NICOLAS ALBERTO MOJICA CRUZ  
YILBER MANUEL MÉNDEZ CALDERÓN

---

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARA  
OPTAR EL TITULO DE:  
INGENIERO EN ENERGÍA

DIRECTOR:  
M. Sc. CARLOS ALIRIO DÍAZ GONZÁLEZ

UNIVERSIDAD AUTONOMA DE BUCARAMANGA  
FACULTAD DE INGENIERÍA  
PROGRAMA DE INGENIERÍA EN ENERGÍA  
BUCARAMANGA

2021

## CONTENIDO

INTRODUCCIÓN.....	1
1. MARCO TEORICO .....	4
1.1. CONTEXTO ZNI .....	4
1.2. PCA.....	5
1.3. CLUSTERING.....	6
1.4. POTENCIALES ENERGETICOS DE BIOMASA .....	8
1.5. DATA WRANGLING.....	9
1.6. METRICAS DE DISTANCIA .....	10
2. RECOLECCION Y TRATAMIENTO DE DATOS.....	12
2.1. FUENTES DE INFORMACIÓN .....	12
2.2. VARIABLES DE LA BASE DE DATOS .....	13
2.3. TRATAMIENTO DE LA BASE DE DATOS .....	15
2.4. BASES DE DATOS OBTENIDAS.....	18
2.5. COMPORTAMIENTO DE LAS ZONAS DURANTE 2019-2020 .....	19
2.6. REPRESENTATIVIDAD DE LA MUESTRA.....	22
2.7. BASE DE DATOS DE BIOMASA.....	22
3. PCA .....	24
4. CLUSTERING .....	31
4.1. PARAMETROS DEL CLUSTER.....	31
4.2. CRITERIOS DE SELECCIÓN DE LOS PARAMETROS DE ENTRADA A LA FUNCIÓN KMEDOIDS .....	32
4.2.1. Número óptimo de clústeres y selección de la métrica de distancia .....	32
4.3. FUNCIÓN <i>EVALCLUSTERS</i> .....	33
4.4. PRESENCIA DE OUTLIERS .....	36
5. POTENCIAL ENERGÉTICO DE BIOMASA .....	39
5.1. PRONOSTICO PARA 2020.....	41
5.2. ASIGNACIÓN DE POTENCIAL ENERGETICO BRUTO PARA CADA ZONA 43	
5.3. POTENCIAL TECNICO Y MARGENES DE BIOMASA .....	43
6. ANALISIS DE RESULTADOS.....	47
6.1. TIPOLOGIAS DE LAS CARACTERISTICAS DE LAS ZNI.....	47

6.2. CALCULO DE INDICADORES INDIVIDUALES.....	52
6.3. CALCULO DE INDICADORES COMPUESTOS.....	53
6.4. INTERPRETACIÓN DEL INDICADOR COMPUESTO PARA EVALUAR LA FACTIBILIDAD DEL SISTEMA DE GASIFICACIÓN DE BIOMASA .....	56
6.5. ANALISIS DE SENSIBILIDAD .....	63
6.6. FACTOR DE CORRECCIÓN DEL INDICADOR COMPUESTO .....	64
CONCLUSIONES.....	69
RECOMENDACIONES .....	71
BIBLIOGRAFIA .....	72

## LISTA DE FIGURAS

Figura 1: caracterización de las ZNI expuesta en el sitio web del IPSE .....	5
Figura 2: energía generada durante el mes de septiembre del 2020 para la localidad de BETE, Media Atrato, Chocó. ....	17
Figura 3: Distribución normal del DAU promedio para cada localidad (histórico, para los 4 periodos). ....	20
Figura 4: Distribución normal del DD promedio para cada localidad (histórico, para los 4 periodos). ....	20
Figura 5: Gráfico de dispersión de la desviación típica del valor promedio del DAU para cada localidad. ....	21
Figura 6: Gráfico de dispersión de la desviación típica del valor promedio del DD para cada localidad. ....	21
Figura 5: Biplot de la componente PC2 vs PC1. ....	27
Figura 6: Diagrama de Pareto, realizado para la determinación del número óptimo de componentes. ....	28
Figura 7: Diagrama de dispersión con todas las localidades, vista tridimensional. ....	28
Figura 8: Diagrama de dispersión PC2 vs PC1 con todas las localidades. ....	29
Figura 9: Diagrama de dispersión PC3 vs PC1 con todas las localidades. ....	29
Figura 10: Diagrama de dispersión PC3 vs PC2 con todas las localidades. ....	30
Figura 11: Gráfico de dispersión, vista PC2 vs PC1, con la asignación de los grupos (k=3). ....	34
Figura 12: Gráfico de dispersión, vista PC2 vs PC1, con la asignación de los grupos (k=4). ....	35
Figura 13: Gráfico de dispersión, vista PC2 vs PC1, con la asignación de los grupos (k=5). ....	35
Figura 14: Gráfico de dispersión, grupos (k=5), de los datos con outliers, vista PC2 vs PC1. ....	36
Figura 15: Histogramas para cada una de las componentes principales seleccionadas, presencia de outliers. ....	37
Figura 16: Histogramas para cada una de las componentes principales seleccionadas, sin outliers. ....	38
Figura 17: Cultivos con potencial energético considerable en Colombia (UPME, IDEAM, COLCIENCIAS, UIS). ....	39
Figura 18: Zonas con mayor margen de biomasa seca (DBM), 25 primeras. ....	45
Figura 19: Composición del margen de biomasa seca (DBM) para las 25 zonas con mayor valor, aporte de cada cultivo. ....	46

Figura 20: Histograma para cada uno de los intervalos definidos para la variable CAPACIDAD de las zonas que conforman el C1.....	48
Figura 21: Comportamiento del nCITB para cada clúster.....	54
Figura 22: Comportamiento del nCIWB para cada clúster.....	55
Figura 23: Comportamiento del nCIDB para cada clúster.....	55
Figura 24: Valor y composición del nCITB para las ZNI del C1.....	56
Figura 25: Valor y composición del nCITB para las ZNI del C2.....	57
Figura 26: Valor y composición del nCITB para las ZNI del C3.....	57
Figura 27: Valor y composición del nCITB para las ZNI del C4.....	58
Figura 28: Valor y composición del nCITB para las ZNI del C5.....	58
Figura 29: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C1, discriminadas por municipio.....	59
Figura 30: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C2, discriminadas por municipio.....	60
Figura 31: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C3, discriminadas por municipio.....	60
Figura 32: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C4, discriminadas por municipio.....	61
Figura 33: Visualización de la ubicación geográfica de las ZNI que conforman cada clúster y viabilidad del sistema de gasificación.....	62
Figura 34: Resultados de los aportes de cada término al indicador global para cada clúster, redefiniendo gamma, coeficiente de ponderación del margen de biomasa.....	64
Figura 35: Resultados del indicador compuesto corregido (nCITBcorr) comparado con los aportes de los déficits y biomasa del indicador base global (nCITB) para cada cluster.....	67

## LISTA DE TABLAS

Tabla 1: estado de prestación del servicio para el periodo de septiembre de 2020.....	17
Tabla 2: Número de localidades para cada uno de los periodos comprendidos en los datos	19
Tabla 3: Valores reales para la localidad de San Antoñito, Buenaventura, Valle del Cauca. .....	26
Tabla 4: Valores normalizados con Z-score para la localidad de San Antoñito, Buenaventura, Valle del Cauca.....	26
Tabla 5: Resultados de los parámetros usados para definir el valor de k.....	34
Tabla 6: Valores del PCI empleados para el cálculo del potencial energético de cada cultivo. .....	40
Tabla 7: Factores de residuo empleados para calcular la generación de residuos de cada cultivo.....	41
Tabla 8: Valores máximos de rendimiento utilizados para el pronóstico del año 2019 y 2020. .....	43
Tabla 9: Mínimos y máximos globales de los valores reales para las variables que caracterizan la prestación de energía.....	47
Tabla 10: Tabla de frecuencias para la variable CAPACIDAD de las zonas que conforman el C1.....	48
Tabla 11: Tipología de las zonas que conforman el clúster 1 (C1).....	49
Tabla 12: Valores reales de la ZNI representativa, medoide del C1.....	49
Tabla 13: Tipología de las zonas que conforman el clúster 2 (C2).....	49
Tabla 14: Valores reales de la ZNI representativa, medoide del C2.....	49
Tabla 15: Tipología de las zonas que conforman el clúster 3 (C3).....	50
Tabla 16: Valores reales de la ZNI representativa, medoide del C3.....	50
Tabla 17: Tipología de las zonas que conforman el clúster 4 (C4).....	50
Tabla 18: Valores reales de la ZNI representativa, medoide del C4.....	51
Tabla 19: Tipología de las zonas que conforman el clúster 5 (C5).....	51
Tabla 20: Valores reales de la ZNI representativa, medoide del C5.....	51
Tabla 21: Marcas de clase, de las variables que caracterizan el servicio de prestación de la energía para cada clúster.....	52
Tabla 22: Valores de los indicadores individuales para la localidad de Limones, del municipio de Guapi.....	65

## ACRONIMOS

ZNI	Zonas no interconectadas
IPSE	Instituto de Planeación y Promoción de Soluciones Energéticas para las ZNI
DAU	Déficit de atención a usuarios
DD	Déficit de diésel
PC	Componente Principal
PCA	Análisis de Componentes Principales
UPME	Unidad de Planeación Minero-Energética
IDEAM	Instituto de Hidrología, Meteorología y Estudios Ambientales
DBM	Dry biomass margin (Margen de biomasa seca)
WBM	Wet biomass margin (Margen de biomasa húmeda)
TBM	Total biomass margin (Margen de biomasa total)
CI	Composite Index (índice compuesto)
PCI	Poder calorífico inferior
SIN	Sistema interconectado nacional
CNM	Centro Nacional de Monitoreo
SUI	Sistema Único de Información de Servicios Públicos Domiciliarios
CAP	Capacidad instalada en kW
USU	Número de usuarios reportados
EFACT	Cantidad de energía facturada al mes kWh/mes
CU	Costo unitario del kWh, COP/kWh
FACT	Cantidad de dinero correspondiente a la energía facturada, COP/mes
SUBS	Cantidad de dinero correspondiente al subsidio, COP/mes
DIE	Cantidad de diésel comprado, en galones (GL)
PDIE	Precio unitario del galón de diésel, COP/GL
PTDIE	Precio del transporte total del diésel, COP
CTDIE	Costo total unitario del diésel, COP/GL
CH	Calinski Harabasz Index
S	Silhouette Coefficient
GapS	Gap Statistic Index
DB	Davies Bouldin Index
k	Número de clústeres
PE	Potencial energético
FR	Factor de recolección de residuos
FC	Factor de corrección para el indicador compuesto

*Si alguna de los acrónimos inicia en “n” indica normalizado. Por ejemplo*

*nCITB: Indicador compuesto con biomasa total normalizado.*

*nDAU: Déficit de atención a usuarios, normalizado.*



## RESUMEN

La prestación del servicio de energía eléctrica para las zonas no interconectadas de Colombia se caracteriza por ser precario en temas de calidad y confiabilidad. Por lo cual, a lo largo del tiempo se han ido planteando diversas soluciones energéticas con el fin de mejorar las condiciones del servicio. Sin embargo, el tipo de solución más común es la energía solar, sin ahondar en otros tipos de generación, en este estudio se orienta a la gasificación de biomasa debido a que la biomasa en el país representa alrededor del 17 por ciento de las fuentes disponibles para producir energía. Para poder determinar la viabilidad de implementar la gasificación fue necesario realizar una reducción de características basada en la metodología de Análisis de Componentes Principales, de la cual, con solo 3 componentes se logra abarcar un 91.03% de la varianza de los datos, con estas se emplea el algoritmo de clusterización kmedoids. Al graficar los resultados de la clusterización se logra observar una conformación de grupos ( $k=5$ ) claramente diferenciados. A partir de los grupos se definieron 5 tipologías de ZNI referentes a sus características de generación y se calcularon indicadores compuestos que expresan si en una zona es o no viable la implementación del sistema de gasificación como solución energética.

*Palabras claves:* ZNI, gasificación de biomasa, clusterización, PCA, kmedoids, indicadores compuestos.

## ABSTRACT

The provision of electric power service in non-interconnected areas of Colombia is characterized by being precarious in terms of quality and reliability. For this reason, several energy solutions have been proposed over time in order to improve service conditions. However, the most common type of solution is solar energy, without delving into other types of generation, this study focuses on biomass gasification because biomass in the country represents about 17 percent of the available sources for energy production. In order to determine the feasibility of implementing gasification, it was necessary to perform a feature reduction based on the Principal Component Analysis methodology, of which, with only 3 components it is possible to cover 91.03% of the variance of the data, with these the kmedoids clustering algorithm is used. When plotting the clustering results, a conformation of clearly differentiated groups ( $k=5$ ) can be observed. From the groups, 5 typologies of ZNI were defined according to their generation characteristics and composite indicators were calculated to express whether or not the implementation of the gasification system as an energy solution is feasible in a given zone.

*Key words:* ZNI, biomass gasification, clustering, PCA, kmedoids, composite indicators.

## INTRODUCCIÓN

El acceso al servicio de energía eléctrica en Colombia se divide en dos grandes partes, las zonas que están conectadas al Sistema Interconectado Nacional (SIN) y las que no, que reciben el nombre de Zonas No Interconectadas (ZNI), estas últimas cubren un 52% del territorio nacional y se caracterizan por encontrarse en lugares de difícil acceso y por contar con riqueza en recursos naturales. Dichas zonas se abastecen de energía en términos muy precarios de calidad del servicio, poca confiabilidad y horas de suministro, a pesar de que cuentan con soluciones de energización locales mediante la generación por fuentes no renovables (diésel mayoritariamente).

El aprovechamiento de los recursos naturales de dichas zonas representa una alternativa a considerar con el fin de ofrecer un servicio constante que permita mejorar la calidad de vida de los usuarios, por esta razón se considera en este estudio la gasificación de la biomasa como solución energética.

Para el gobierno de Colombia ha sido de vital importancia que los servicios básicos sean llevados a cada uno de los rincones del país, por lo cual, a este le interesa que la prestación del servicio de energía eléctrica sea monitoreada en dichas zonas. Por consiguiente, se tiene como entidad encargada al *Centro Nacional de Monitorio (CNM)* que hace parte del *Instituto de Planeación y Promoción de Soluciones Energéticas para las ZNI (IPSE)*. Este organismo emplea el uso de telemetría, tecnología que permite la medición remota de magnitudes físicas como la velocidad, temperatura, voltaje y más específicamente energía, potencia, horas de suministro y posterior envío de la información (Muñoz, 2003). Sin embargo, la cantidad de zonas que cuentan con telemetría son relativamente pocas y están en las cabeceras municipales principalmente. Para el resto de las zonas que no cuentan con dicha tecnología la recolección de la información se hace a partir de medios electrónicos, como correo, o llamadas telefónicas. Incluso por medio de encuestas realizadas ya sea al prestador de servicio o a los usuarios.

En la conformación de la base de datos, es común encontrar que algunas zonas no cuentan con la información referente a la prestación del servicio, esto debido posiblemente a que no hubo comunicación entre el *CNM* y el prestador del servicio. A pesar de que exista esa situación no es considerable descartar dichas localidades. Entonces, la metodología utilizada se basa en reemplazar dichos valores faltantes por los reportados en otros periodos de tiempo para la misma localidad. De esta manera, es posible obtener una base de datos más completa por tanto, es más representativa.

El análisis de componentes principales se realiza con el fin de poder representar las características de la prestación del servicio de las localidades en un solo gráfico, que puede llegar a contar con máximos 3 componentes. Puesto que, solo es posible visualizar hasta 3 dimensiones. Esto permite poder evidenciar de manera gráfica como están conformados las futuras agrupaciones que son obtenidas a partir del proceso de clusterización. Para este último proceso es necesario definir una serie de parámetros como bien lo son el número clústeres a obtener.

La realización de la metodología de componentes principales y la de clustering se realiza mediante la interfaz de Matlab, en la cual se realiza el código que permite hallar las componentes principales, para la segunda metodología, el clustering, se emplean las funciones disponibles en la interfaz, entre las cuales se tienen clúster jerárquicos y particionales, la variante empleada es el k-medoids que hace parte de los algoritmos de partición. Esta fue empleada debido a que asigna como representante de los grupos una zona en específico, es decir, un individuo existente en la data, a diferencia de por ejemplo el k-means, en la cual, el representante puede ser un individuo no existente.

A partir de las agrupaciones, se analizan las características de cada uno de los individuos que conforman dichos grupos y con esto, es posible definir tipologías de cada uno de los clústeres. Estas representan a los individuos que conforman cada grupo. Considerando las tipologías predeterminadas se orientan estos resultados para evaluar la factibilidad de implementar la gasificación de la biomasa como solución energética. Se orienta a partir de una base de datos de biomasa extraída de las *Evaluaciones Agropecuarias Municipales*, archivo en donde se reportan las hectáreas de cultivo que es producido en los municipios a nivel nacional, con esto y en conjunto con los factores de generación de residuos reportados en el *Atlas de Biomasa Residual* en Colombia se calculan los potenciales energéticos disponibles en cada localidad mediante la asignación de la biomasa del municipio por medio de una ponderación en función del número de usuarios.

Finalmente, la viabilidad se define a partir de una serie de indicadores compuestos resultantes de una expresión que cuenta con indicadores como el déficit de atención a usuarios (*DAU*), el déficit de diésel (*DD*) y los márgenes de biomasa tanto de biomasa seca, húmeda o total. Se realiza un análisis de sensibilidad con el fin de observar que tanto varía el comportamiento de los indicadores compuestos de las ZNI a diferentes criterios de ponderación de sus indicadores individuales.

## **OBJETIVO GENERAL**

Realizar clusterización de zonas no interconectadas de Colombia orientada al uso de la gasificación de biomasa como fuente energética.

## **OBJETIVOS ESPECIFICOS**

- Obtener un arreglo matricial de las “n” y “m” características de las zonas no interconectadas con telemetría y sin telemetría de Colombia.
- Reducir el número de columnas del arreglo matricial de las ZNI por medio de un algoritmo en MATLAB basado en el método de Principal Component Analysis (PCA).
- Realizar un agrupamiento o clusterización de las “n” zonas no interconectadas por medio de un algoritmo en MATLAB.

- Comparar tipologías de zonas no interconectadas de acuerdo con los parámetros seleccionados ( suministro de energía, calidad, disponibilidad del recurso, generación de energía, entre otras).

La estructura de este documento se realiza con el fin de demostrar el cumplimiento de los objetivos, por lo cual se definen cinco ítems o fases de la presente investigación:

- Almacenamiento y tratamiento de los datos: recopilación de los datos de operación y datos técnicos correspondientes a la generación de energía en las ZNI. Así mismo, la obtención de datos agrícolas para la futura obtención del potencial energético de biomasa. Filtrado de ZNI teniendo en cuenta una serie de criterios que no favorecen la generación de energía con gasificación de biomasa. Manejo de datos faltantes (NaN).
- PCA: determinación del número y de las variables óptimas para garantizar una varianza explicada considerable como resultado del PCA.
- Clustering: determinación de los parámetros óptimos para la realización de los clústeres por medio del algoritmo de kmedoids.
- Potencial energético de biomasa: cálculo de los márgenes de energía a obtener a partir de biomasa húmeda y seca para cada ZNI.
- Análisis de resultados: observación de los resultados obtenidos, definición de las tipologías de cada grupo según las características de las ZNI que lo conforman. Formulación de indicadores compuestos para definir la viabilidad de implementar gasificación de biomasa en las ZNI.

Con esto se quiere construir una base teórica que permita de cierto modo evaluar la factibilidad del uso de la gasificación de la biomasa como fuente energética en la mayoría de las ZNI brindando de esta manera un resultado calculado netamente a partir de datos de gobierno, es decir evaluando la factibilidad según datos de generación, capacidad instalada, facturación, subsidios, y características del diésel (precio, cantidad y valor del transporte) y no teniendo en cuenta datos demográficos o culturales de la población.

Así mismo, la generación con biomasa aprovecha los grandes recursos naturales que tiene nuestro país en cuestión de tierra fértil, esto ha demostrado que incentiva la economía de estas zonas debido a que un sistema de gasificación requiere de procesos previos como la recolección del material orgánico, secado y entre otros que son procesos que requieren de personal.

## 1. MARCO TEORICO

Para la elaboración del presente estudio fue necesario abarcar los siguientes elementos teóricos:

### 1.1.CONTEXTO ZNI

Las zonas no interconectadas de Colombia (ZNI) son los municipios, corregimientos, localidades, y caseríos no conectados al Sistema Interconectado Nacional (SIN) Como ejemplo de ZNI encontramos el Archipiélago de San Andrés, Providencia y Santa Catalina, Leticia en el Amazonas, Puerto Carreño en el Vichada y Mitú en Vaupés (Sistema Único de Información Normativa, 2003). Sin embargo hay que aclarar que las ZNI que superan los 500 kW de capacidad instalada no fueron tenidas en cuenta en el presente estudio puesto que, para esos valores de capacidad se vuelve inviable la generación a partir de gasificación de biomasa por el tamaño del gasificador, mayor dificultad en la implementación del sistema y debido a que los costos suben drásticamente, por ejemplo, para un sistema de gasificación basado en un gasificador downdraft el costo de capital ronda los 500-1000 euros por kWe instalado, a diferencia de los basados en gasificadores de lecho fluidizado que está alrededor de los 10000 y 15000 euros por kWe (Bocci, y otros, 2014), empleando el *Chemical Engineering Plant Cost Index (CEPCI)* es posible actualizar estos costos a 2019, valor del índice más reciente conocido, pues están reportados en el 2014. El CEPCI a 2014 es 576 (Turton , Shaeiwitz, Bhattacharyya, & Whiting, 2018) y el 2019 es 607.5 (Jenkins, 2020). Con esto, para un sistema con gasificador downdraft el costo está entre 528-1055 euros por kWe mientras que para el sistema con gasificador de lecho fluidizado está entre 10547-15820 euros por kWe.

Estas se caracterizan por su baja densidad poblacional, por estar ubicadas a una larga distancia de los centros urbanos, por la dificultad de acceso y por su gran riqueza de recursos naturales. Representan el 52% del territorio nacional pero por las razones anteriores resulta muy costoso integrarlas al SIN y se hace necesario que la prestación del servicio se genere directamente en cada zona, en general, las ZNI cuentan con una capacidad operativa que ronda los 280 MW de los cuales solo el 5% provienen de fuentes no convencionales de energía y renovables a pesar de que cuentan con abundantes recursos naturales (Superintendencia de Servicios Públicos Domiciliarios, 2020).

El Instituto de Planeación y Promoción de Soluciones Energéticas para las Zonas No Interconectadas de Colombia (IPSE) en relación con el Sistema Único de Información de Servicios Públicos Domiciliarios (SUI) reportan que las ZNI hacen presencia en 20 de los 32 departamentos que conforman la división política de Colombia, de los cuales prestan servicios en 74 municipios registrando 1702 localidades, información dada a 2019. El término localidad es utilizado para identificar de forma general a un lugar que puede ser una cabecera municipal, un centro poblado o una zona en donde el servicio de energía eléctrica no se presta por medio del SIN.

Actualmente existe una tipificación de las ZNI y esta se realizó según lo establecido en el artículo 6° de la Resolución MME 182138 de 2007, en la cual estas tipologías se definen a

partir del número de usuarios, siendo las de tipo 1 las localidades que cuentan con un número de usuarios mayor a 300; las de tipo 2 entre 151 usuarios y 300; tipo 3 entre 51 y 150 y finalmente las de tipo 4 que cuentan con máximo 50 **usuarios o suscriptores**. De las 1702 reportadas en 2019 se conoce que el 87.5% de las localidades cuentan con máximo 150 usuarios (48.5% zonas tipo 3 y 39% zonas tipo 4) y el restante las conforman zonas de tipo 2 y 1 (10.2% y 2.4% respectivamente). No obstante solo las zonas de tipo 1 abarcan un 42.1% de los usuarios totales.

Los usuarios totales de esas 1702 localidades son 193.871 de los cuales su estratificación está compuesta por un 80.44% de usuarios estrato 1, seguida de un 8.8% y 4.76% de estratos 2 y 3 respectivamente, las demás zonas que equivalen alrededor de un 6% están dadas por estratos 4, 5 y 6 contando también con los usuarios oficiales y comerciales, siendo estos últimos un 4.02%.

Actualmente, en el sitio web del IPSE presentan la caracterización de las ZNI en Colombia de la siguiente manera:

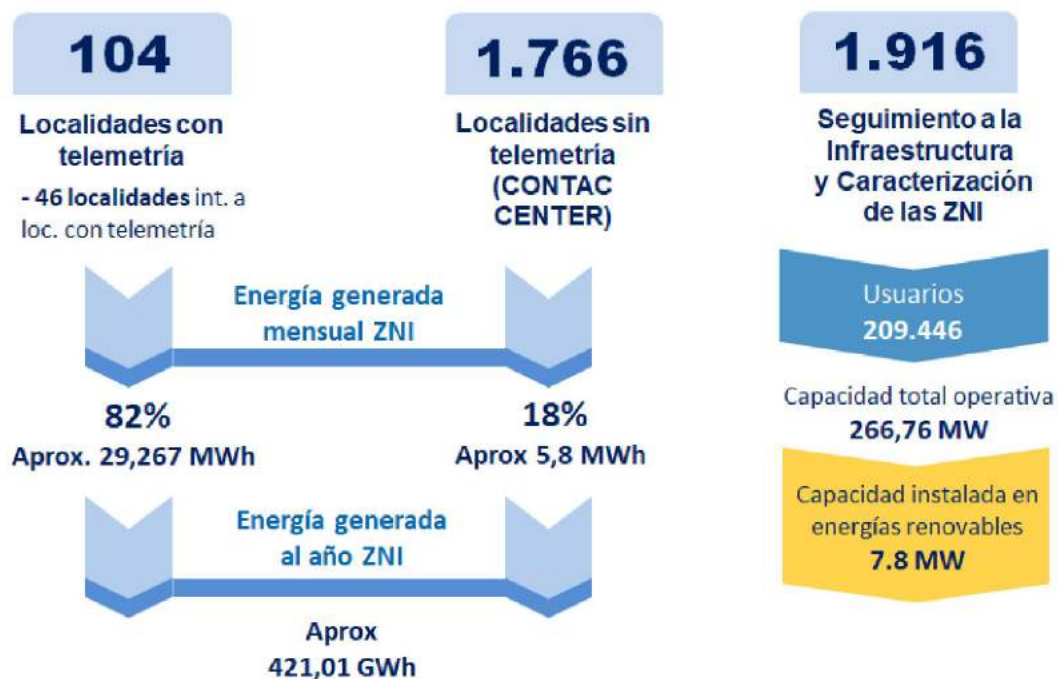


Figura 1: caracterización de las ZNI expuesta en el sitio web del IPSE

### 1.2.PCA

A razón de la necesidad de poder representar las características de las ZNI como un punto en específico en un diagrama de dispersión se implementa el *Análisis de Componentes Principales* (PCA por sus siglas en inglés, *Principal Component Analysis*). De manera practica se puede afirmar que el análisis de componentes principales transforma el conjunto

de variables originales en un conjunto más pequeño de variables, las cuales son combinaciones lineales de las primeras, que contienen la mayor parte de la variabilidad presente en el conjunto inicial (Peña, 2002). Por consiguiente, PCA también permite facilitar la interpretación de la información contenida en los datos.

El análisis por componentes principales tiene como propósito central la determinación de unos pocos factores (componentes principales) que retengan la mayor variabilidad contenida en los datos. Las nuevas variables poseen algunas características estadísticas “deseables”, tales como independencia y no correlación. Puesto que, una variable correlacionada linealmente no aporta más información a un conjunto de datos.

Para obtener las componentes principales PCA no ignora las covarianzas o correlaciones pero se concentra en las varianzas, a tal punto, que la función objetivo es encontrar la combinación lineal de las variables originales que maximice la varianza. Lo anterior se puede ejemplificar con la siguiente expresión:

$$PC_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1p}x_p = \sum_{j=1}^p \alpha_{1j}x_j \quad (1.1)$$

Donde:

$\alpha_{ij}$ : es el loading o peso de la variable  $x_j$  para la componente principal  $PC_i$ .

A partir de la primer PC se obtiene la segunda que cuenta como característica que es ortogonal a la primera, es decir, no se encuentran correlacionadas entre sí. Adicionalmente, que maximiza la varianza, este proceso se repite hasta obtener la misma cantidad de variables iniciales si se quiere, sin embargo, como el objetivo es reducir la dimensionalidad, lo normal es que se busquen un número de variables menor, en este caso menor a p.

### 1.3.CLUSTERING

Al igual que con el PCA el clustering parte de un objetivo o necesidad, este puede ser definido como: *Dado un conjunto de puntos de datos, dividirlos en un conjunto de grupos en los cuales sus individuos sean lo más parecidos posible.* (University of Minnesota, Department of Computer Science and Engineering, 2014) A lo largo del tiempo, algunos usos habituales en los que se plantea el problema de agrupación son los siguientes:

- En algunas situaciones se emplea como paso previo antes de implementar otra metodología referente a la analítica de datos, esto puesto que, el clustering se puede considerar como una forma de resumen de datos o para el análisis de datos atípicos.
- Como filtro, al conocer las agrupaciones se pueden identificar las características de estos y por consiguiente, es posible descartar o seleccionar alguno de ellos para un estudio más profundo.
- Para resumir los datos, debido a que la clusterización está ampliamente relacionada con los métodos de reducción dimensional, por lo cual, la integración de estas

metodologías puede ser útil al crear representaciones más compactas y fáciles de interpretar en un amplia variedad de aplicaciones.

Los algoritmos de clustering más estudiados son el particional y el jerárquico. Estos algoritmos han sido muy utilizados en una amplia gama de aplicaciones debido principalmente a su simplicidad y facilidad de implementación en relación con otros algoritmos. Los particionales tienen como objetivo descubrir las agrupaciones presentes en los datos optimizando una función objetivo específica y mejorando iterativamente la calidad de los datos. Estos algoritmos suelen requerir ciertos parámetros del usuario para elegir los puntos prototipo que representan cada cluster.

Por otro lado, los algoritmos de agrupación jerárquica abordan el problema de la agrupación mediante desarrollando una estructura de datos basada en un árbol binario llamado dendrograma. Una vez construido el dendrograma, se puede elegir automáticamente el número correcto de conglomerados dividiendo el árbol en diferentes niveles para obtener diferentes soluciones de clustering para el mismo conjunto de datos sin tener que volver a ejecutar el algoritmo de clustering. A diferencia del cluster particional en el cual pueden producirse conjuntos de resultados totalmente diferentes dependiendo del criterio utilizado durante el proceso.

Los métodos particionales deben contar con un conjunto de clusters iniciales que luego se mejoran de forma iterativa. Los métodos jerárquicos, en cambio, pueden empezar con los puntos de datos individuales y construir el clustering. El papel de la métrica de distancia también es diferente en ambos algoritmos. En el clustering jerárquico, esta se aplica inicialmente a los puntos de datos en el nivel base y luego se aplica progresivamente a los subconglomerados. Sin embargo, en el caso de los métodos particionales, en general, los puntos representativos elegidos en diferentes iteraciones pueden ser puntos virtuales como el centroide del clúster (que no existe en los datos).

La obtención de un clúster o agrupación a partir de una algoritmo particional se resume:

---

**Algoritmo: Clúster particional**

---

**1:** *Seleccionar los  $K$  individuos que serán los puntos usados como centroides.*

**2:** *Inicia el ciclo. **Repetir.***

**3:** *Formar los  $K$  grupos asignando cada punto al centroide más cercano.*

**4:** *Recalcular el centroide para cada clúster.*

**5:** *El criterio de convergencia es encontrado. **Hasta.***

Los criterios de convergencia comúnmente se asocian a la iteración en la cual ninguno de los miembros del cluster fue reasignando, es decir, la configuración anterior es idéntica a la actual. También se toma como criterio de convergencia estadísticos como la *Suma de Errores Cuadrados* (SSE, por sus siglas en ingles *Sum of Squared Errors*) o la *Suma Residual de*



*Cuadrados* (RSS, por sus siglas en inglés *Residual Sum of Squares*). La formulación matemática de las anteriores funciones objetivo se muestra a continuación:

Dado un set de datos  $D = \{x_1, x_2, \dots, x_N\}$  que contiene N individuos y que al aplicar o asignar unos clusters, los centroides están dados por  $C = \{C_1, C_2, \dots, C_K\}$ . El objetivo es minimizar SSE, por consiguiente está se define como:

$$SSE(C) = \sum_{k=1}^K \sum_{x_1 \in C_k} (x_i - c_k)^2 \quad (1.2)$$

Donde  $c_k$  es el centroide de  $C_K$ , este ultimo representa el número de individuos que conforman el clúster K.

$$c_k = \frac{\sum_{x_1 \in C_k} x_1}{|C_K|} \quad (1.3)$$

#### 1.4.POTENCIALES ENERGETICOS DE BIOMASA

La biomasa es la materia orgánica utilizada como fuente energética. Por su amplia definición, esta abarca un extenso conjunto de materias orgánicas que se caracteriza por su heterogeneidad, tanto por su origen como por su naturaleza. Colombia por ser un país rico en biodiversidad se considera uno de los países con mayores potenciales, a tal punto que la biomasa en el país representa alrededor del 17 por ciento de las fuentes disponibles para producir energía (FINAGRO, 2021). Esto representa una oportunidad para prestar el servicio de electricidad en las ZNI. Ambientalmente, reduce las emisiones de gases de efecto invernadero y también ayuda a los países a mantener los compromisos internacionales en materia de cambio climático, específicamente a Colombia en lo propuesto en el COP21.

El potencial energético de la biomasa se define como la cantidad de energía almacenada en una cierta cantidad de materia, generalmente se suele asociar con el poder calorífico inferior (PCI) puesto que, se expresan en las mismas unidades (energía / masa). Los potenciales energéticos de la biomasa se calculan de diferente manera según el tipo de sector en donde se produce:

- En el sector agrícola, se busca obtener el potencial referente a los cultivos y el modelo matemático se expresa en función del área cultivada, la generación del producto principal y la cantidad de residuos resultantes. Este tipo de biomasa generalmente se aprovecha mediante procesos termoquímicos. Inicialmente el modelo establece la masa seca del residuo agrícola, por unidad de tiempo, se puede expresar como:

$$M_{rs} = A * R_c * M_{rg} * Y_{rs} \quad (1.4)$$

Donde:

$M_{rs}$ : Masa de residuo seco [tonelada/año]

$A$ : Área cultivada [hectárea/año]

$R_c$ : Rendimiento del cultivo [tonelada producto / hectárea sembrada]

$M_{rg}$ : Masa de residuo generada del cultivo [tonelada residuo / tonelada de producto]

$Y_{rs}$ : Fracción de residuo seco [tonelada residuo seco / tonelada de residuo húmedo]

Una vez obtenida la masa del residuo de biomasa agrícola seca, el potencial energético de la biomasa residual agrícola (PEBR), conformada por los residuos de cosecha, se evalúa mediante la multiplicación de la masa por el  $PCI$ .

- Para el sector pecuario, la alternativa difundida para el aprovechamiento de la energía contenida en el estiércol de bovinos, porcinos y aves de jaula se orienta hacia la digestión anaerobia, mientras que para el estiércol de aves con cama (engorde), cuyo contenido de humedad es menor al 60%, la tecnología aplicable es la termoquímica. Por lo cual se debe emplear dos modelos para cada caso:

El modelo de digestión anaerobia se aplica en los caso de bovinos, porcinos y aves ponedoras, debido a que la presencia de altos niveles de humedad y de microorganismos que facilitan la producción de metano. Mientras que el modelo térmico se emplea debido a que las aves de engorde por las condiciones de su hábitat producen un estiércol con menor contenido de humedad (Murillo, 1999) No obstante, no se detalla en el cálculo de estos modelos debido a que no se tiene en cuenta este tipo de biomasa en el presente estudio para las ZNI.

### 1.5.DATA WRANGLING

Los conjuntos de datos tienen frecuentemente datos faltantes, duplicados, erróneos o incluso en algunas ocasiones están mal estructurados, encasillados en columnas que no corresponden. Por lo cual se define el data Wrangling como el proceso de recopilar, seleccionar y transformar los datos para emplearlos en procesos analíticos. También se conoce como limpieza de datos o “munging”. Este proceso se conoce por ser tedioso al tener que examinar los datos y a pesar de que hay herramientas tecnológicas desarrolladas y que evolucionan constantemente, entender y preparar un conjunto de datos recién adquirido para su uso posterior requiere de mucho tiempo y esfuerzo, alrededor de entre un 50% y 80% (Endel & Piringer, 2015). Este proceso se realiza con el fin de obtener calidad en los datos, por lo cual se hace necesario emplear diversas fuentes de información.

El procesamiento de los datos se da generalmente a través de una serie de etapas:

- La adquisición describe cómo obtener acceso a los datos, ya sea a través de bases de datos disponibles o generando datos nuevos, por ejemplo, a través de una encuesta u observaciones.
- En la etapa de extracción, los datos se convierten desde cualquier formato de entrada que se haya adquirido (por ejemplo, archivos XLS, PDF o incluso documentos de texto sin formato) en un formato que se pueda utilizar para su posterior procesamiento y análisis.

- Limpiar y transformar los datos a menudo implica eliminar registros no válidos y traducir todas las columnas. utilizar un conjunto sensato de valores. También puede combinar dos conjuntos de datos diferentes en una sola tabla, eliminar duplicados entradas o aplicar cualquier número de otras normalizaciones. A medida que adquiera datos, notará que dichos datos a menudo tienen muchas inconsistencias. Este es un proceso que siempre debe realizarse.
- Análisis de datos para responder preguntas particulares como en nuestro caso, que se tiene como objetivo establecer tipologías de las zonas a partir de sus características (variables relacionadas a la generación de energía).
- Presentar la data, proceso que se realiza para ejemplificar el tratamiento que se le ha empleado a los conjuntos de datos.

## 1.6. METRICAS DE DISTANCIA

Una métrica de distancia es una función que define una distancia entre dos observaciones. Algunas métricas de distancia comunes son: la distancia euclidiana, distancia euclidiana estandarizada, distancia cityblock, distancia Mahalanobis, distancia de Minkowski, distancia de Chebychev, distancia del coseno (cosine), distancia de correlación, distancia Hamming, distancia Jaccard y distancia de Spearman.

Las distancias entre un vector  $X_s$  y uno  $X_t$ , se definen de la siguiente manera:

- Distancia Minkowski

$$d_{st} = \sqrt[p]{\sum_{j=1}^n |x_{sj} - x_{tj}|^p} \quad (1.5)$$

Para el caso especial de  $p = 1$ , la distancia de Minkowski da la distancia cityblock. Para el caso especial de  $p = 2$ , la distancia de Minkowski da la distancia euclidiana. Para el caso especial de  $p = \infty$ , la distancia de Minkowski da la distancia Chebychev. Como se observa en las próximas definiciones.

- Distancia euclidiana

$$d_{st}^2 = (x_s - x_t)(x_s - x_t)' \quad (1.6)$$

La distancia euclidiana es un caso especial de la distancia Minkowski, donde  $p = 2$ .

- Distancia de Chebychev

$$d_{st} = \max_j \{|x_{sj} - x_{tj}|\} \quad (1.7)$$

La distancia de Chebychev es un caso especial de la distancia Minkowski, donde  $p = \infty$ .

- Distancia cityblock

$$d_{st} = \sum_{j=1}^n |x_{sj} - x_{tj}| \quad (1.8)$$

La distancia cityblock es un caso especial de la distancia Minkowski, donde  $p = 1$ .

- Distancia Mahalanobis

$$d_{st}^2 = (x_s - x_t)C^{-1}(x_s - x_t)' \quad (1.9)$$

Donde C es la matriz de covarianza.

- Distancia euclidiana estandarizada

$$d_{st}^2 = (x_s - x_t)V^{-1}(x_s - x_t)' \quad (1.10)$$

Donde V es un vector de factores de escala.

- Distancia coseno (cosine distance)

$$d_{st} = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}} \quad (1.11)$$

- Distancia de correlación (correlation distance)

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}} \quad (1.12)$$

Donde

$$\bar{x}_s = \frac{1}{n} \sum_j x_{sj} \quad \text{y} \quad \bar{x}_t = \frac{1}{n} \sum_j x_{tj}$$

De manera similar se formulan la distancia de Jaccard y spearman.

## 2. RECOLECCION Y TRATAMIENTO DE DATOS

En esta sección se detallan las fuentes de las cuales se realizó la recopilación de los datos, así como también las diversas acciones implementadas para el tratamiento de los mismos.

### 2.1. FUENTES DE INFORMACIÓN

Las ZNI hacen presencia en 20 de los 32 departamentos que conforman la división política de Colombia, de los cuales prestan servicios en 74 municipios registrando 1702 localidades, información dada a 2019. Para la realización de esta tesis se recopilieron datos de diferentes fuentes, ordenadas de mayor a menor relevancia:

- *Sistema Único de Información de Servicios Públicos Domiciliarios (SUI)*: es el sistema oficial del sector de servicios públicos domiciliarios del país que recoge, almacena, procesa y publica información reportada por parte de las empresas prestadoras y entidades territoriales. De este organismo se emplearon la mayoría de los datos, debido a que estos fueron entregados en formatos específicos, presentando uniformidad en la información. Los formatos empleados son: *Formato ZNI TO1* y el *Formato ZNI C5*. El primero entrega los registros de operación diarios y mensuales de las ZNI en donde reportan variables como las horas de servicio, los usuarios, la energía facturada, los subsidios, la capacidad instalada, entre otras. Mientras que el segundo formato brinda información referente a las características técnicas de generación como son los galones comprados, el precio del galón, el costo del transporte del combustible, punto de acopio, etc. Estos formatos muestran el cumplimiento de los prestadores de servicio referente a la Circular SSPD-IPSE No. 0001 de 2004 y la Circular SSPD-IPSE No. 124 de 2008.
- *Centro Nacional de Monitoreo (CNM) departamento que hace parte del Instituto de Planificación y Promoción de Soluciones Energéticas para las ZNI (IPSE)*: este departamento es el encargado de realizar actividades fundamentales para el seguimiento a la prestación del servicio de energía eléctrica en las ZNI por tanto recopila y suministra información oportuna de los parámetros eléctricos de dichas zonas. Sin embargo, el formato de estos datos a pesar de que son libres no facilita la implementación en el estudio, puesto que, se muestran en formato PDF en informes mensuales, a diferencia de los formatos del SUI que son directamente archivos XLS o CSV con los datos en bruto. No obstante, la mayoría de la información concuerda entre ambas fuentes.
- *Datos Abiertos de Colombia*: sitio web en la cual es posible observar, descargar y usar datos del gobierno colombiano para investigaciones u otras aplicaciones. Según la fuente, cuenta con datos procedentes de 933 entidades, entre las cuales se encuentra el IPSE, en este caso fueron empleados los archivos identificados como: *“Estado de la prestación del servicio de energía en Zonas No Interconectadas”*. Estos archivos también es posibles obtenerlos en formato CSV.

## 2.2. VARIABLES DE LA BASE DE DATOS

A partir de las fuentes anteriormente mencionadas se logró obtener una base de datos compuesta de las siguientes variables:

1. **Departamento:** obtenida de fuentes.
2. **Municipio:** obtenida de fuentes.
3. **Nombre de la localidad:** obtenida de fuentes.
4. **Tipo de cabecera:** valor correspondiente a la localidad definida en el campo 3, ya sea, un centro poblado, localidad menor o cabecera municipal. Obtenida de fuentes.
5. **Nombre del prestador de servicio:** nombre de la empresa que presta el servicio de energía a la localidad definida en el campo 3. Obtenida de fuentes.
6. **Longitud:** valor de la coordenada geográfica para la localidad definida en el campo 3. Obtenida de fuentes.
7. **Latitud:** valor de la coordenada geográfica para la localidad definida en el campo 3. Obtenida de fuentes.
8. **Tipo de localidad:** tipificación según lo establecido en el artículo 6° de la Resolución MME 182138 de 2007, en la cual estas tipologías se definen a partir del número de usuarios, siendo las de tipo 1 las localidades que cuentan con un número de usuarios mayor a 300; las de tipo 2 entre 151 usuarios y 300; tipo 3 entre 51 y 150 y finalmente las de tipo 4 que cuentan con máximo 50 usuarios o suscriptores.
9. **Capacidad instalada:** corresponde a la potencia instalada del o los grupos electrógenos en la localidad definida en el campo 3. Obtenida de fuentes.
10. **Potencia promedio:** valor promedio de potencia utilizada durante el mes en cuestión. Calculada a partir de la energía facturada y las horas de servicio.
11. **Potencia máxima:** valor máximo de potencia utilizada durante el mes en cuestión. Calculada a partir de la energía promedio.
12. **Potencia de diseño:** valor de potencia necesario para que la localidad definida en el campo 3 opere nominalmente.
13. **Factor de carga:**
14. **Horas de servicio:** cantidad de horas diarias promedio con energía eléctrica. Obtenida de fuentes.
15. **Rango de servicio:** intervalo de horas de servicio. Obtenida de fuentes.
16. **Usuarios E1:** cantidad de usuarios pertenecientes al estrato 1. Obtenida de fuentes.
17. **Usuarios totales:** cantidad de usuarios totales de la localidad definida en el campo 3. Obtenida de fuentes.
18. **Energía facturada E1:** cantidad de energía facturada por los usuarios pertenecientes al estrato 1. Obtenida de fuentes.
19. **Energía total:** cantidad de energía total facturada. Obtenida de fuentes.
20. **kWh/usuario:** cantidad de energía entregada para cada usuario perteneciente a la localidad definida en el campo 3. Calculada.
21. **kWh/usuario diseño:** cantidad de energía a entregar a cada usuario de la localidad definida en el campo 3 asignada según el consumo de subsistencia para cada tipo de localidad (campo 8). Obtenida de fuentes.

22. **Energía facturada E1 de diseño:** cantidad de energía total para cumplir con el consumo de subsistencia de los usuarios pertenecientes al estrato 1 de la localidad definida en el campo 3.
23. **Energía total de diseño:** cantidad de energía total para cumplir con el consumo de subsistencia de los usuarios pertenecientes a la localidad definida en el campo 3.
24. **Usuarios E1 atendidos:** cantidad de usuarios realmente atendidos si se le entrega la energía que define el consumo de subsistencia según el tipo de localidad. Calculada.
25. **Déficit de atención usuarios:** porcentaje de déficit (valor negativo) o superávit (valor positivo), de la atención de usuarios. Si es negativo significa que se está satisfaciendo la necesidades de energía eléctrica de menos usuarios de los que realmente hay. Si es positivo, se están atendiendo más de los reales. Calculada.
26. **Facturación:** cantidad de dinero cobrado por el prestador del servicio referente a la energía eléctrica suministrada a todos los usuarios de la localidad definida en el campo 3. Obtenida de fuentes.
27. **Subsidio:** cantidad de dinero recibido por el prestador del servicio para subsidiar parte de la energía eléctrica suministrada a todos los usuarios de la localidad definida en el campo 3. Obtenida de fuentes.
28. **Costo unitario:** valor del kWh en la localidad definida en el campo 3. Obtenida de fuentes.
29. **Diesel comprado:** cantidad de galones de diésel adquiridos para el sistema de generación de la localidad definida en el campo 3. Obtenida de fuentes.
30. **Precio por galón:** valor unitario del galón de diésel. Obtenida de fuentes.
31. **Costo del transporte de diésel:** valor total de transportar todos los galones de diésel comprados, desde el punto de acopio a la localidad definida en el campo 3. Obtenida de fuentes.
32. **Costo total por galón:** valor unitario total de un galón de diésel puesto en la localidad definida en el campo 3. Obtenida de fuentes.
33. **Punto de acopio:** lugar o zona donde el diésel llega para ser enviado a cada localidad en cuestión. Punto de distribución del combustible. Obtenida de fuentes.
34. **Déficit de diésel:** porcentaje de déficit (valor negativo) o superávit (valor positivo), de la cantidad de diésel. Si es negativo significa que se cuenta con menos diésel del necesario, si es positivo indica lo contrario. Calculada.
35. **Energía comprada:** cantidad de energía disponible en el total de galones de diésel comprados. Calculada.
36. **Combustible requerido:** cantidad de combustible necesario para suplir con el consumo de subsistencia de todos los usuarios. Calculada.
37. **Eficiencia eléctrica:** expresada por la relación entre la energía facturada y la energía contenida en el diésel. Calculada.
38. **Eficiencia real:** valor de la eficiencia del grupo electrógeno calculada a partir del factor de carga. Calculada.
39. **Costo de transporte por galón:** valor de transportar un galón desde el punto de acopio hasta la localidad definida en el campo 3. Calculada.

**40. Galones por kW:** relación entre la cantidad de galones comprados y la potencia instalada en la localidad, expresa la disponibilidad de combustible por cada kW instalado.

De las 40 columnas que conforman la base de datos, 23 de ellas fueron obtenidas de las diversas fuentes ya mencionadas y 17 de ellas calculadas a partir de las anteriores. Cada una de ellas fue tomada en cuenta para la realización de diferentes etapas durante el proyecto.

### **2.3. TRATAMIENTO DE LA BASE DE DATOS**

El almacenamiento de la información por parte del CNM se basa en dos tipos de sistemas para hacer el seguimiento a la prestación del suministro de energía eléctrica. Por un lado, se tiene la telemetría o teledatada que permite la medición remota de la energía eléctrica y el posterior envío de esta información hacia el Centro Nacional de Monitoreo. Por otra parte, en las localidades sin telemetría se realiza el seguimiento mediante un contacto telefónico, correo electrónico u otros medios de comunicación digital (Contac Center), bajo la modalidad entrevistas o encuestas a los usuarios y/o operadores locales. En cualquiera de los dos casos existe la probabilidad de que sean reportados datos erróneos, ya sea por fallas del sistema de telemetría o por errores humanos. Por consiguiente, fue necesario revisar la data con el fin de encontrar posibles errores y así mismo tratar con los valores perdidos.

Generalmente las variables categóricas (Departamento, municipio, nombre de localidad, entre otras) no presentaron problemas con valores perdidos. No obstante, las variables numéricas como la capacidad instalada, la energía facturada, las horas de servicio, la cantidad de galones, la facturación, el subsidio, entre otras si presentaron dicho problema. A continuación se muestran las acciones realizadas para tratar el problema de los datos faltantes para cada variable que los presentó:

- Capacidad instalada [kW]: se combinaron las 3 fuentes mencionadas en el inciso 2.1. sin importar el mes del reporte, puesto que se parte de la consideración que la capacidad instalada de las ZNI normalmente no cambia, no se agregan ni se pierden los grupos electrógenos.
- Horas de servicio: se combinaron las 3 fuentes mencionadas en el inciso 2.1. teniendo en cuenta el mes del cual se está construyendo la base de datos, se tuvo como consideración un mes antes y uno después. Si el caso persiste, se opta por asignar un valor basado en zonas que hacen parte del mismo municipio y cuentan con valores similares de capacidad y número de usuarios.
- Usuarios: el número de usuarios se pudo asignar solo mediante las 3 fuentes mencionadas en el inciso 2.1. De este caso no se obtuvieron grandes casos, este es un dato que es reportado con gran facilidad.
- Energía facturada: se asignó solo mediante las 3 fuentes mencionadas en el inciso 2.1. Mismo caso que con los usuarios. Sin embargo, si lo anterior no fue suficiente y se contaba con la capacidad instalada y el número de horas de servicio, entonces se calculó de la siguiente manera:



$$E_{fact} = Cap_{ins} * h_{serv} * 30 \left[ \frac{kWh}{mes} \right] \quad (2.1)$$

Donde:

$E_{fact}$ : es la energía facturada por los prestadores de servicio para cada ZNI.

$Cap_{ins}$ : capacidad instalada en kW.

$h_{serv}$ : horas de servicio.

- Facturación: se combinaron las 3 fuentes mencionadas en el inciso 2.1. Si el problema persiste, y se cuenta con el costo unitario del kWh, se calcula con la expresión:

$$Fact = E_{fact} * CU \left[ \frac{COP}{mes} \right] \quad (2.2)$$

Donde:

$Fact$ : es la cantidad de dinero facturada por la prestación del servicio de energía eléctrica para cada ZNI.

$E_{fact}$ : es la energía facturada por los prestadores de servicio para cada ZNI.

$CU$ : es el costo unitario del kWh.

- Costo unitario: se combinaron las 3 fuentes mencionadas en el inciso 2.1. Si el problema persiste, y si es posible se calcula a partir de despejar la expresión anterior, si no es posible matemáticamente, se emplea asignar el valor del  $CU$  a partir de comparar localidades del mismo municipio que reciban el servicio de energía eléctrica a partir del mismo prestador del servicio, ya que, el  $CU$  no varía con ese criterio. También, en algunos de los formatos se cuenta con la distribución del  $CU$ , lo correspondiente a generación, distribución, transmisión, entre otros y con la suma se obtiene el  $CU$ .
- Cantidad de diésel comprado: este valor depende de cada zona, por consiguiente, solo se optó por buscar en las 3 fuentes mencionadas en el inciso 2.1.
- Precio por galón: el precio por cada galón depende más que todo del lugar, por tanto, para los valores perdidos fue asignado a partir de comparar las zonas del mismo municipio y que cuenten con el mismo punto de acopio.
- Costo del transporte: al igual que con el parámetro anterior se realizó teniendo en cuenta el punto de acopio y el municipio. En algunos municipios se encontró una correlación lineal, es decir, el costo total de transportar el combustible crece linealmente.

Es necesario tener en cuenta que no se generen referencias circulares en la base de datos. Por ejemplo, tratar de calcular la facturación a partir de la energía facturada y el  $CU$  y también emplear esta misma instrucción para calcular el  $CU$ . Cabe aclarar que si las anteriores acciones no permitieron obtener los datos de alguna ZNI en cuestión, esta fue eliminada del

estudio por falta de información, esta situación es relativamente común en estos lugares a tal punto que en los informes de prestación del servicio se reportan. Para septiembre de 2020 el CNM reportó:

Estado	Número de localidades
Con servicio	1608
Servicio intermitente	6
Sin servicio	87
Sin información	71
<b>Total</b>	<b>1772</b>

Tabla 1: estado de prestación del servicio para el periodo de septiembre de 2020.

Solucionado el tema de los datos faltantes, aún es necesario detallar lo realizado para los datos atípicos, entre los cuales se observaron zonas con mayor generación de energía que la cantidad de energía disponible en el diésel, así como también zonas en las que la facturación no coincide con la energía facturada, es decir, el valor reportado no coincide con el resultado de la expresión 2.2. Adicionalmente, zonas en las cuales la eficiencia eléctrica es superior a la eficiencia de un motor de combustión interna. Finalmente, también se consideraron como datos atípicos las zonas que presentaron servicio intermitente o no contaron con el servicio de energía en ese mes, esto se observó a partir de los informes de prestación de servicio del CNM, mediante las siguientes gráficas:

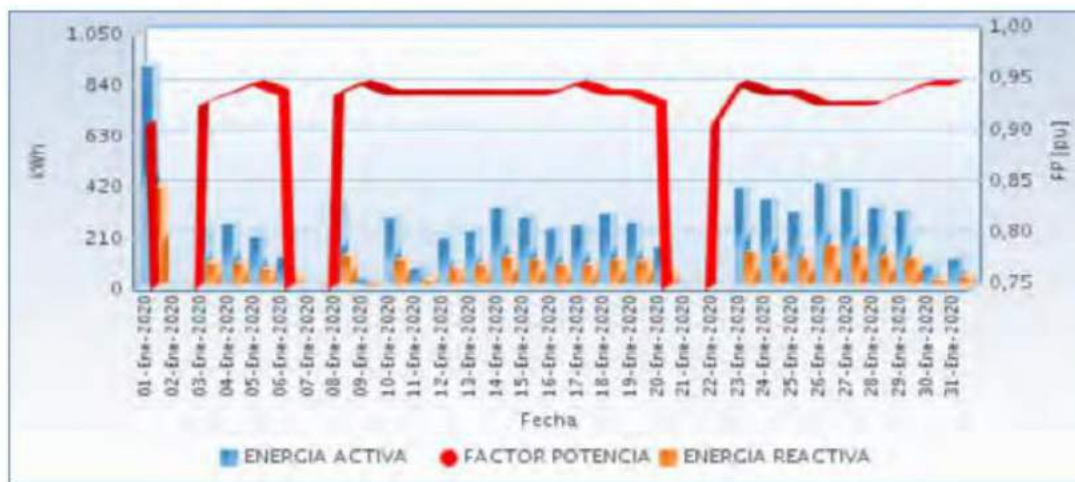


Figura 2: energía generada durante el mes de septiembre del 2020 para la localidad de BETE, Media Atrato, Chocó.

Como se observa la gráfica para la localidad de BETE, del municipio de Medio Atrato, Choco esta zona presentó servicio intermitente, pues en los días del 2, 7, 21 y 22 no tuvo servicio de energía eléctrica y adicionalmente durante los demás días fue inconsistente en la generación. Generando durante el mes de septiembre un total de energía de 7833.45 kWh por lo cual este valor debió ser corregido tomando en cuenta un el comportamiento de la generación durante agosto y octubre, que son los meses antes y después. Por tanto, recibió un valor de 15582 que representa de mejor manera la generación de energía en esta zona durante un mes “normal”.

De la misma manera que con la energía eléctrica las horas de servicio promedio reportadas se ven afectadas por estos casos de servicio intermitente o sin prestación de servicio. Por consiguiente, lo realizado se toma a partir de la energía corregida, se calculan las horas de servicio promedio diarias, por medio de la expresión 2.1. A su vez, en dichos casos fue también necesario corroborar los rangos de servicio, el rango de servicio en el cual se generó energía.

Con el fin de corroborar la validez de las variables como la cantidad de galones comprados se calculó la eficiencia eléctrica con la que está operando la localidad. Con este parámetro se encontró zonas con muy baja eficiencia menor al 10% y zonas que superan el 50%. Los anteriores son valores atípicos en este parámetro, por tanto, se corrigieron a partir de observar los reportes de meses cercanos para comprobar al cambiar el dato que tanto variaba la eficiencia, hasta el punto de que se encontrará en valores más coherentes. Si la eficiencia era muy alta, se buscaba una cantidad de galones que hayan sido comprados en otro mes más grande al actual, así la eficiencia bajaría, o lo contrario en el otro caso.

#### **2.4. BASES DE DATOS OBTENIDAS**

Al momento de realizar esta tesis el periodo más reciente reportado fue el de septiembre del 2020. Por consiguiente, se planteó abarcar 4 periodos de tiempo con el fin de considerar que tanto cambia el comportamiento de dichas zonas, por lo cual, se trabajó con un rango de 6 meses para atrás. Es decir, se conformaron 4 bases de datos, correspondientes a los meses de marzo y septiembre de los años 2019 y 2020 respectivamente. El proceso de almacenamiento y tratamiento de datos para cada uno de los periodos fue el mismo que se detalla en los ítems anteriores. Por otra parte, a continuación se describen algunos criterios con los cuales algunas zonas no fueron tenidas en cuenta:

- Las zonas en las cuales el prestador de servicio implementa una fuente diferente a la generación mediante diésel, entre éstas se tienen sistemas híbridos entre diésel y fotovoltaica, diésel y biomasa o micro centrales hidroeléctricas y diésel. De igual manera, las zonas que generan solo a partir de otra fuente, como solo fotovoltaica. Ejemplos: la localidad de isla fuerte, del departamento de Bolívar cuenta con generación a partir de diésel y fotovoltaica. Por este criterio fueron descartadas 49 zonas, en su mayoría hacen parte del departamento del Choco y Vaupés.
- Las zonas cuya capacidad instalada supera los 500 kW, puesto que, la implementación de un sistema de gasificación de biomasa superior a esa capacidad instalada deja de ser viable, debido a que el tipo de gasificador downdraft deja de ser una opción, por lo cual se debe implementar un tamaño más grande de gasificador lo que también provoca que se deba tener un sistema de tratamiento y limpieza para los gases más profundo. Entre estos destacan San Andrés y Providencia, Leticia en Amazonas e Inírida en Vaupés, que son cabeceras municipales. Las anteriores mencionadas y alrededor de otras 30 más.
- Las zonas que no fue posible completar los datos.

Aplicando lo anterior mencionado, las bases de datos cuentan con la siguiente cantidad de zonas:

Periodo de la base de datos	Número de localidades
Septiembre de 2020	1589
Marzo de 2020	1607
Septiembre de 2019	1606
Marzo de 2019	1552

Tabla 2: Número de localidades para cada uno de los periodos comprendidos en los datos

## 2.5. COMPORTAMIENTO DE LAS ZONAS DURANTE 2019-2020

Para el estudio resulta importante observar el comportamiento de las ZNI durante el tiempo, puesto que, la clusterización se realiza a partir de los valores de las variables numéricas, los grupos se obtienen a partir de la similitud matemática de los individuos que conforman cada uno de los conjuntos. Por tanto, si los individuos presentan mucha variación implicaría que la clusterización se realizara de manera independiente para cada periodo, porque hasta las escalas podrían variar.

Por lo cual fue necesario realizar un análisis que determine que tanta difiere el comportamiento de las variables asociadas a la generación y que son reportadas por el SUI y el CNM durante los periodos de tiempo estudiados. Para esto se parte de dos parámetros calculados:

- El déficit de atención a usuarios (DAU): este parámetro relaciona de manera directa el número de usuarios y la energía generada.
- El déficit de diésel (DD): se calcula a partir de la energía generada y los galones comprados.

Estos parámetros entre si están influenciados por las demás variables tenidas en cuenta en el estudio. Puesto que, estos parámetros representan las causas por las cuales la calidad del servicio de energía eléctrica en estos lugares es precaria, generalmente reportan fallas en los sistemas de generación o disponibilidad de combustible deficiente. Esto se da posiblemente por temas económicos como pueden ser los precios del combustible o del transporte de este, o por otro lado, por aspectos que definen el sistema de generación (capacidad, las horas de servicio, entre otros).

Como se observa en la tabla 2, las ZNI son diferentes en cada uno de los periodos de tiempo, por lo cual, se optó por mezclar dichas bases de datos y observar cómo varían los dos parámetros mencionados durante los 4 periodos de tiempo para las zonas que están presentes en todos. De la mezcla se obtuvo un total de 1632 localidades de las cuales 1397 están presentes en todos los periodos de tiempo.

El análisis constó de calcular el promedio del *DAU* y el *DD* de cada localidad, al igual que la desviación estándar, con esto se obtiene el promedio y desviación resultante de los valores para cada momento de tiempo. Teniendo la media para cada zona se calcula la distribución normal obteniendo:

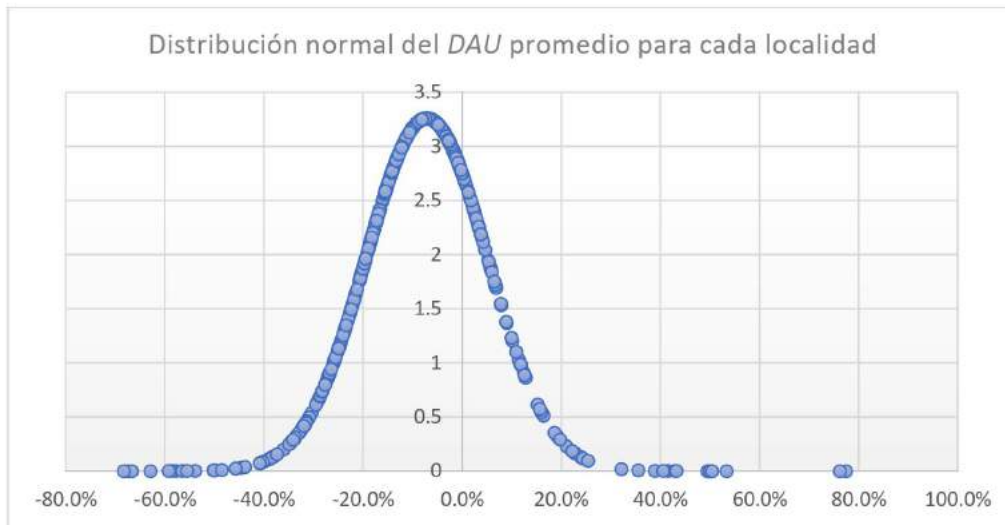


Figura 3: Distribución normal del DAU promedio para cada localidad (histórico, para los 4 periodos).

De la figura anterior, es posible afirmar que la mayoría de las zonas mantienen un *DAU* negativo es decir, que están supliendo las necesidades de menos usuarios de los reales. Esta problemática se logra evidenciar a partir de la misma gráfica pero para el *DD*:

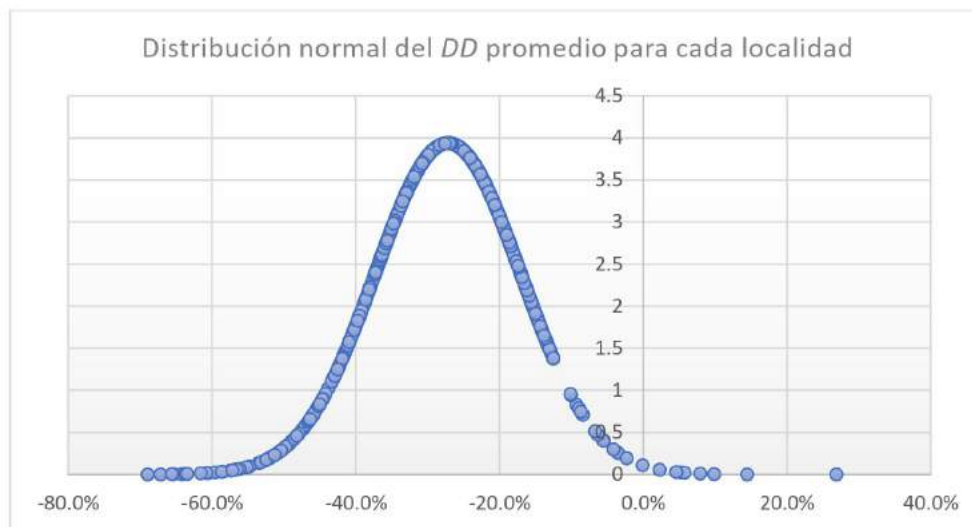


Figura 4: Distribución normal del DD promedio para cada localidad (histórico, para los 4 periodos).

Claramente se observa que solo unas cuantas localidades cuenta con superávit de diésel, sin embargo, también es posible afirmar que a pesar de que muchas de las localidades cuentan con *DD* negativo, parece ser, que algunas de ellas lo administran de mejor manera, pues se encuentra menos orientada a la izquierda la distribución normal del *DAU*. Para concluir si el comportamiento de las *ZNI* cambia drásticamente se realizó un gráfico de dispersión que muestra la desviación estándar vs la media para cada uno de los parámetros:

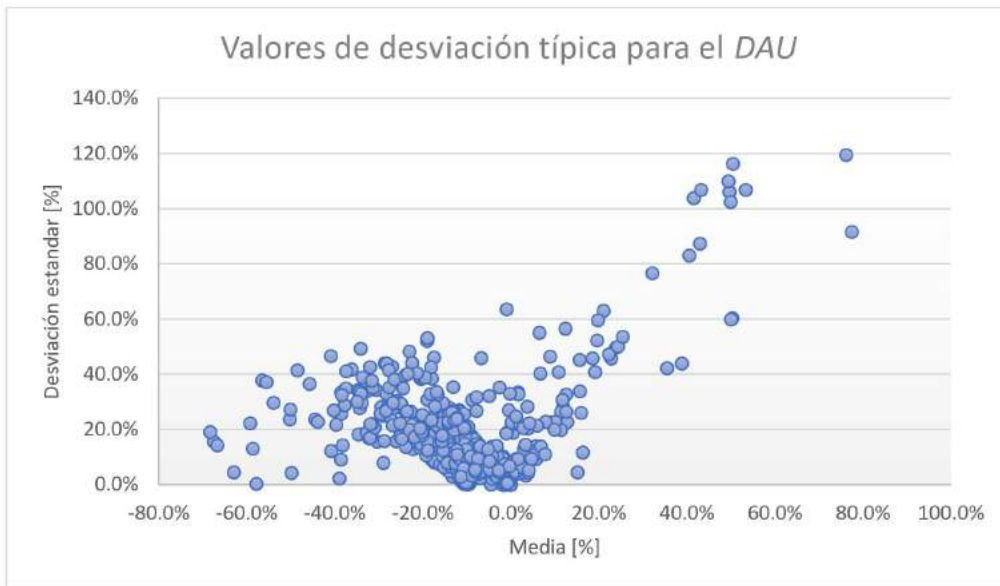


Figura 5: Gráfico de dispersión de la desviación típica del valor promedio del DAU para cada localidad.

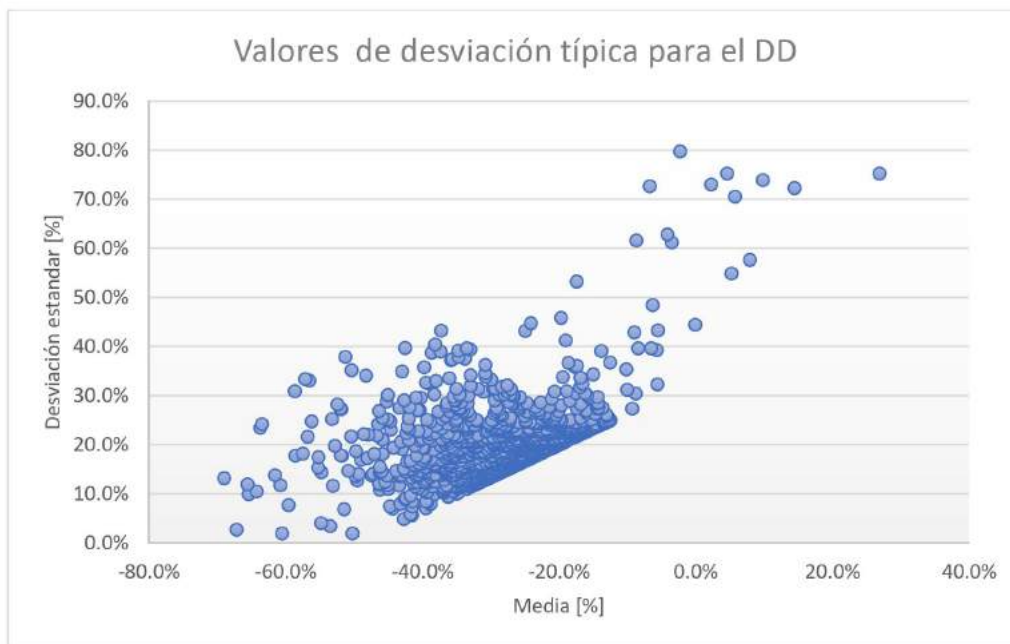


Figura 6: Gráfico de dispersión de la desviación típica del valor promedio del DD para cada localidad.

Las gráficas anteriores muestran que la desviación estándar porcentual para la media del DD y DAU para la mayoría de las zonas no supera el 30% del valor del parámetro. Adicionalmente, es posible observar que las zonas que presentan mayor variación son las que están en los extremos, es decir, las que tienen un déficit más positivo o negativo. A tal punto que en esos valores se tienen márgenes de desviación del 100% es decir, más o menos el valor promedio del déficit en cuestión. Por las razones anteriormente mencionadas se

concluye que la variación es importante solo para aquellas zonas que presentan valores altos de los dos parámetros ya sean positivos o negativos, pero al ser la gran mayoría ubicada entre valores de -0.3 a 0.1 en el caso del *DAU* y de -0.4 a -0.1 para el *DD* se concluye que la variación no es considerable para tener que realizar los procedimientos detallados en las siguientes secciones para todos los 4 periodos de tiempo.

## 2.6. REPRESENTATIVIDAD DE LA MUESTRA

A partir del ítem anterior fue posible demostrar que las variables que caracterizan la prestación de servicio de energía eléctrica en las localidades no varían de gran forma con el pasar del tiempo. Sin embargo, se emplea este otro inciso con el fin de calcular el porcentaje de error que se asume al tomar una muestra de 1397 zonas del total que se tienen reportadas 1632. Para eso se parte de la ecuación que determina el tamaño de una muestra para una población finita:

$$n = \frac{N * Z^2 * p * q}{e^2 * (N - 1) + Z^2 * p * q} \quad (2.3)$$

Donde:

$n$ : tamaño de la muestra buscado

$N$ : tamaño de la población

$p$ : probabilidad de que ocurra el evento estudiado

$q = 1 - p$ : probabilidad de que no ocurra el evento estudiado

$e$ : error de estimación máximo aceptado

$Z$ : parámetro estadístico que depende del nivel de confianza

En este caso, no se quiere hallar el tamaño de la muestra sino el error máximo al que se está sometiendo el estudio al trabajar con una muestra de 1397 ZNI. Para esto, es necesario definir el resto de los parámetros involucrados en la ecuación. Entonces, al desconocer la probabilidad de que ocurra el evento, se asigna un 50% a "p" y un 50% a "q", es lo que comúnmente se realiza al igual que asumir un nivel de confianza del 95% para el cual el parámetro  $Z$  tiene un valor de 1.96 (Software para encuestas, QuestionPro, 2020).

Con los datos anteriores y aplicando establecer objetivo mediante el *Solver* de Excel, se obtuvo que con ese tamaño de muestra se está trabajando con un error máximo de 0.995 %. Por lo cual, la muestra se considera representativa debido a que en múltiples referencias se considera un error máximo del 5%.

## 2.7. BASE DE DATOS DE BIOMASA

Adicionalmente a la base de datos que caracteriza la prestación del servicio en las ZNI es necesario obtener una referente a la biomasa disponible en las mismas. Para este objetivo se encontró una base histórica de los años 2007 a 2018, relacionada con la producción agrícola

Nacional. Esta base de datos está disponible en *Datos Abiertos de Colombia* fuente número 3 definida en el ítem 2.1. Se encuentra con el nombre de: “*Evaluaciones Agropecuarias Municipales EVA*” y son datos proporcionados por el Ministerio de Agricultura y Desarrollo. Cuenta con columnas de interés como el área sembrada, el área cosechada, la producción y el rendimiento, y adicionalmente define si el cultivo es permanente o transitorio.

Para esta base de datos no fue necesario una limpieza de datos como tal sino más bien una reestructuración para poder ser empleados en el estudio. Este proceso se verá detallado en la sección 5: *potencial energético de la biomasa*.



### 3. PCA

En esta sección se describe el procedimiento necesario para implementar la metodología de reducción de dimensionalidad y los criterios empleados en el desarrollo de esta.

Como ya se sabe de la anterior sección se obtuvo una matriz que cuenta con 1589 zonas para el periodo de *septiembre 2020* de las cuales se tienen 49 columnas compuestas de variables categóricas y numéricas. Como estas variables tienen diferentes rangos y escalas, es necesario centralizar y estandarizar, debido a que el proceso de PCA identifica aquellas direcciones en la que la varianza es mayor y si no se realiza estas acciones, las variables cuya escala sea mayor dominarán al resto. La centralización y estandarización está dada por la siguiente expresión:

$$Dc_{ij} = \frac{D_{ij} - \bar{D}_j}{\sigma(D_j)} \quad (3.1)$$

Donde:

$Dc_{ij}$  es el valor centralizado de la variable  $j$  para la ZNI  $i$ .

$D_{ij}$  es el valor real de la variable  $j$  para la ZNI  $i$ .

$\bar{D}_j$  es el promedio de los valores reales de la variable  $j$ .

$\sigma(D_j)$  es la desviación estándar de la variable  $j$ .

La expresión anterior también recibe de normalización Z-score, en esta estrategia un valor idéntico a la media recibe el valor de cero, si es menor obtendrá un valor negativo y si es mayor uno positivo.

Posterior a la normalización Z-score se procede a calcular la matriz de covarianzas de los datos normalizados, de la cual se extraen los *eigenvalues* y *eigenvectors*. Estos últimos se definen de la siguiente manera: “*Son todos aquellos vectores que, al multiplicarlos por una matriz, el resultado es el mismo vector o un múltiplo entero del mismo*”. Por otro lado, a los eigenvalues se les conoce como: “*El valor por el que se multiplica un eigenvector*”. Los eigenvectors cumplen con una serie de propiedades matemáticas:

- Los eigenvectors solo existen para matrices cuadradas y no para todas. En el caso de que una matriz  $n \times n$  tenga eigenvectors, el número de ellos es  $n$ .
- Si se escala un eigenvector antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo eigenvector. Esto se debe a que si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud pero la dirección es la misma.
- Todos los eigenvectors de una matriz son perpendiculares (ortogonales) entre ellos, independientemente de las dimensiones que tengan.

En el PCA se aprovecha la segunda propiedad mencionada y por tanto, todos los eigenvectors que se obtienen se escalan de tal forma que su magnitud sea 1 y de este modo se consigue

que todos sigan un estándar. A cada componente principal se le asigna un eigenvector y el orden de las componentes se establece por orden decreciente de eigenvalue. Así pues, la primera componente es el eigenvector con el eigenvalue asociado más alto.

Como ya se mencionó en el marco teórico las componentes principales son una combinación lineal de los variables, estos coeficientes que multiplican cada una de las variables se les conoce como *loadings* e indican el peso o la importancia que tiene dicha variable en el cálculo de la componente. Estos loadings están almacenados en cada una de las posiciones de los eigenvectors. Por tanto, el valor de las PC's para cada ZNI se calcula al multiplicar la matriz de datos normalizados por los eigenvectors. Con esa última acción se concluye el proceso de reducción de dimensionalidad, sin embargo, es necesario evaluar el número de componentes óptimos y para eso se emplea lo que se denomina proporción de la varianza explicada, la varianza explicada es la varianza total de los datos, y la varianza es la desviación estándar al cuadrado.

Los eigenvalues a su vez representan la varianza de cada componente, por consiguiente, la proporción de varianza explicada puede expresarse como:

$$\omega_i = \frac{Eigenvalue_i}{\sum Eigenvalues} \quad (3.2)$$

Donde:

$\omega_i$  es la proporción de la varianza explicada para la componente principal  $i$ .

La suma de este último parámetro es un fundamento propicio para definir el número óptimo de componentes, sin embargo, cabe aclarar que en nuestro caso se busca obtener un número máximo de 3 componentes principales debido a que solo es posible graficar 3 dimensiones.

Se realizó este procedimiento con la matriz de datos mencionada al inicio de esta sección correspondiente al periodo de septiembre de 2020 con esta se logró observar que la proporción de varianza explicada de las 3 primeras componentes no logra superar el 60% de la varianza total, mientras que un valor óptimo debe ser del 80% o superior (Gorgas & Cardiel, 2011). Por tanto, se trabajó realizando múltiples configuraciones que se diferenciaban en el número de variables, de lo anterior se observó que al reducir el número de variables era posible perder menos información de los datos, es decir, la proporción de varianza explicada acumulada en las 3 primeras componentes incrementaba.

Adicionalmente, se probó solo ingresando variables que no fueron calculadas a partir de las variables reportadas en las ZNI, es decir, variables como la potencia promedio fueron calculadas a partir de tres variables reportadas en los informes de generación de las ZNI las cuales son la capacidad, las horas de servicio promedio y la energía facturada al mes. Debido a que estas variables calculadas de cierta manera cuentan con información de las variables base, por lo cual tenerlas en cuenta se considera redundante.

A continuación se muestra la configuración final de la matriz de datos de entrada con la que se realizó el PCA. La matriz de las 1589 ZNI contó con las siguientes variables:

- Capacidad expresada en kW
- Número de usuarios
- Energía facturada mensual expresada en  $\frac{kWh}{mes}$
- Costo unitario del kWh expresada en  $\frac{COP}{kWh}$
- Facturación mensual expresada en COP
- Subsidio mensual expresada en COP
- Cantidad de diesel comprado expresada en GL
- Precio del diesel expresada en  $\frac{COP}{GL}$
- Precio del transporte de diesel expresada en COP

Para ejemplificar los procedimientos realizados durante esta y las siguientes secciones se empleará una de las ZNI. Esta es la localidad de San Antoñito (yurumanguí) del municipio de Buenaventura que pertenece al departamento del Valle del Cauca. Esta es un centro poblado y tiene las siguientes características:

CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
50	61	2745	1461,97	4.037.647	3.380.110	1152	6602	593.004

Tabla 3: Valores reales para la localidad de San Antoñito, Buenaventura, Valle del Cauca.

Al normalizar con Z-score se obtiene:

CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
-0.1001	-0.2463	-0.2046	-0.1555	-0.2112	-0.2013	1.6266	-0.3898	0.1741

Tabla 4: Valores normalizados con Z-score para la localidad de San Antoñito, Buenaventura, Valle del Cauca.

La matriz de covarianzas es una matriz cuadrada de nxn en donde n coincide con el número de características, por tanto se extraen nueve eigenvectors y eigenvalues. El eigenvector que corresponde a la primera componente es:

$$\vec{V}_1 = [0.3268 \quad 0.3795 \quad 0.3936 \quad -0.0391 \quad 0.3912 \quad 0.3884 \quad 0.3815 \quad -0.046 \quad 0.3757]$$

Este eigenvector almacena los loadings o pesos de cada una de las 9 variables empleadas en el PCA, y de esta manera es posible expresar el PC1 de la siguiente manera:

$$PC1 = 0.3268CAP + 0.3795USU + 0.3936EFACT - 0.0391CU + 0.3912SUBS + 0.3884DIE + 0.3815DIE - 0.046PDIE + 0.3757PTDIE$$

Por lo cual, el valor del PC1 para la zona de ejemplo es:

$$PC1 = 0.3268 * (-0.1001) + 0.3795 * (-0.2463) + 0.3936 * (-0.2046) - 0.0391 * (-0.1555) + 0.3912 * (-0.2112) + 0.3884 * (-0.2013) + 0.3815 * (1.6266) - 0.046 * (-0.3898) + 0.3757 * (0.1741)$$

$$PC1 = 0.342447$$

Este procedimiento se realiza con cada una de las zonas y con cada uno de los componentes, a partir de los eigenvectores se construyen las combinaciones lineales de las PC's con la

estructura de la expresión (1.1) y con las características normalizadas de cada zona se hace el respectivo cálculo. El cálculo de las componentes permite expresar con gran precisión la información dada por las variables iniciales, de esta forma se realiza la reducción dimensional.

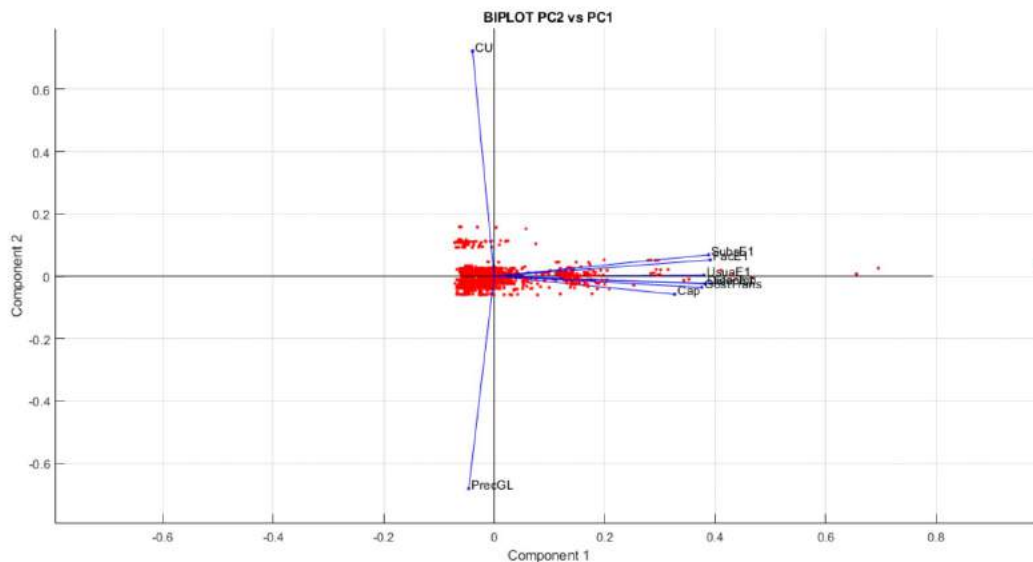


Figura 5: Biplot de la componente PC2 vs PC1.

La figura anterior recibe el nombre de *Biplot* y muestra la influencia que tiene cada una de las variables en torno a las componentes principales, es decir, que este caso en específico el PC1 es ampliamente dominado por la capacidad, los usuarios, la facturación, el subsidio, la cantidad de diésel comprado, el costo del transporte y la energía facturada. En cambio, el PC2 está dominado ampliamente por el CU y el precio del galón. Adicionalmente, es posible afirmar que las variables que influyen en el PC1 son perpendiculares a las que influyen mayormente en el PC2. Adicionalmente, en el caso de las variables influyentes del PC2 se observa que son inversamente proporcionales entre si (CU y PrecGL).

Por otro lado, con los eigenvalues es posible calcular la proporción de varianza explicada que aporta cada componente.

Los eigenvalues obtenidos de la matriz de covarianza de la data son:

$$\overrightarrow{Ev_1} = [6.1733 \ 1.2073 \ 0.8123 \ 0.3824 \ 0.2239 \ 0.1271 \ 0.0609 \ 0.0121 \ 5.74E - 4]$$

Si se suman estas magnitudes se obtiene como resultado nueve, que corresponde a la varianza de las nueve variables ingresadas, puesto que, al normalizar con Z-score se garantiza que tengan media cero y desviación estándar uno, y la varianza al ser el cuadrado de la desviación estándar pues corresponde con esa sumatoria. Por lo cual, el PC1 representa un 68.59% de la varianza total, el PC2 un 13.41% y finalmente el PC3 un 9.02%. De lo anterior se demuestra que con solamente 3 componentes se logra abarcar un 91.03% de toda la varianza de los datos. Esto también se observa con la siguiente figura:

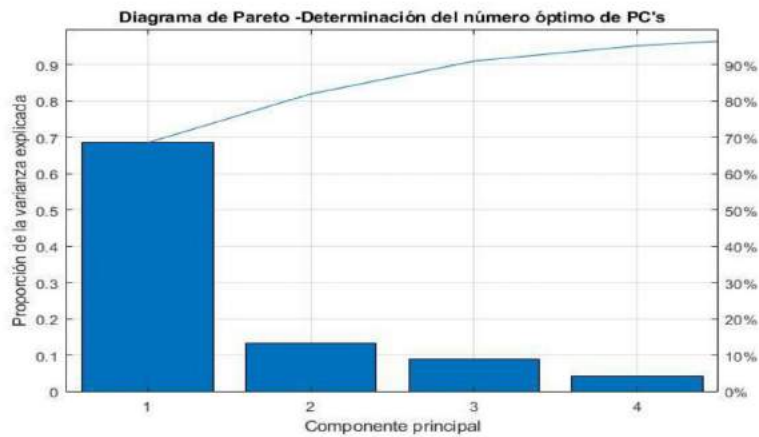


Figura 6: Diagrama de Pareto, realizado para la determinación del número óptimo de componentes.

El diagrama de Pareto muestra el número de componentes hasta que se alcance un 95% de la proporción de varianza acumulada, si se tiene en cuenta la cuarta componente que contribuye con un 4.24% se supera dicho valor, por consiguiente, la contribución de las demás componentes no se encuentra graficada. Con esta representación se concluye la metodología de PCA demostrando que se logró el objetivo de reducir la dimensionalidad de los datos y de esta manera, es posible mostrar cada una de las ZNI como un punto en un diagrama de dispersión garantizando que dicho punto representa con precisión cada una de las características de generación de estas, como se muestra a continuación:

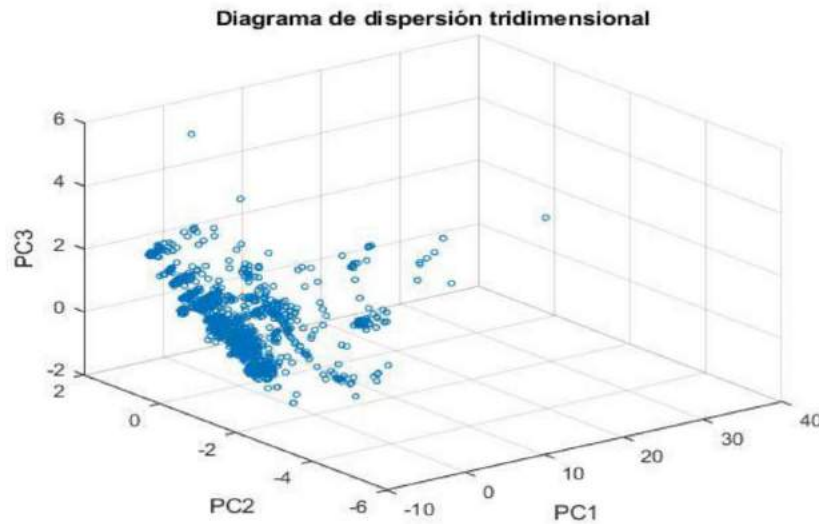


Figura 7: Diagrama de dispersión con todas las localidades, vista tridimensional.

Con la figura 7 es posible observar la información de las 9 variables en una vista de 3 dimensiones, con una precisión de más del 90% como ya se mencionó anteriormente, esto permite poder definir cada ZNI como un individuo en los diagramas de dispersión pudiendo establecer rangos de valores de cada PC y posibles agrupaciones visuales.

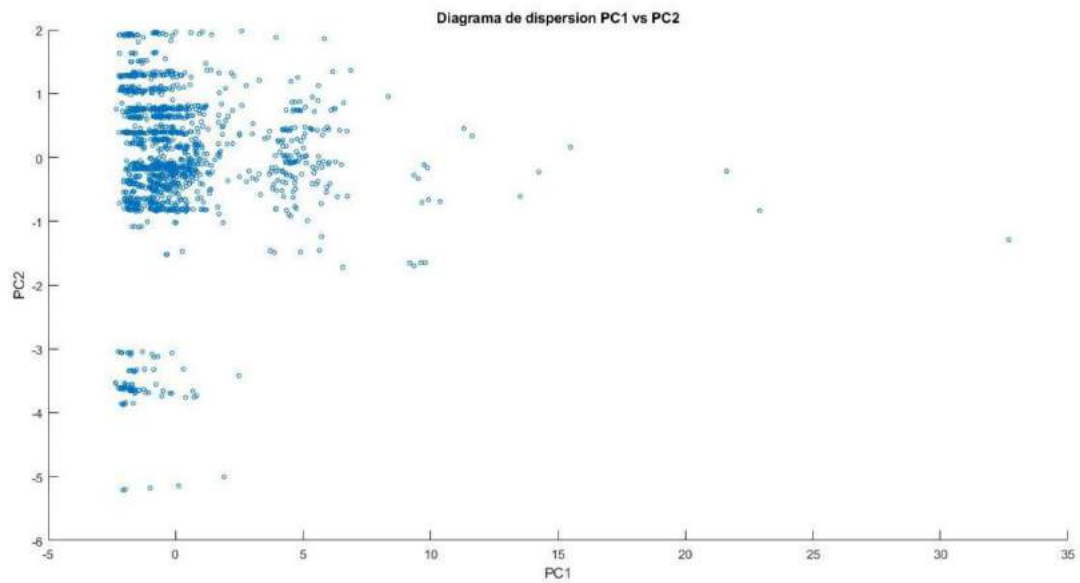


Figura 8: Diagrama de dispersión PC2 vs PC1 con todas las localidades.

Las ZNI en el PC1 están dispersas entre -2 a 35, sin embargo, es posible ver que la mayoría de estas están conformadas en un rango más estrecho -2 a 10 aproximadamente, esta situación se abarca con mayor amplitud más adelante. Por otro lado, en el PC2 se observa que las ZNI varían entre -5 a 2.

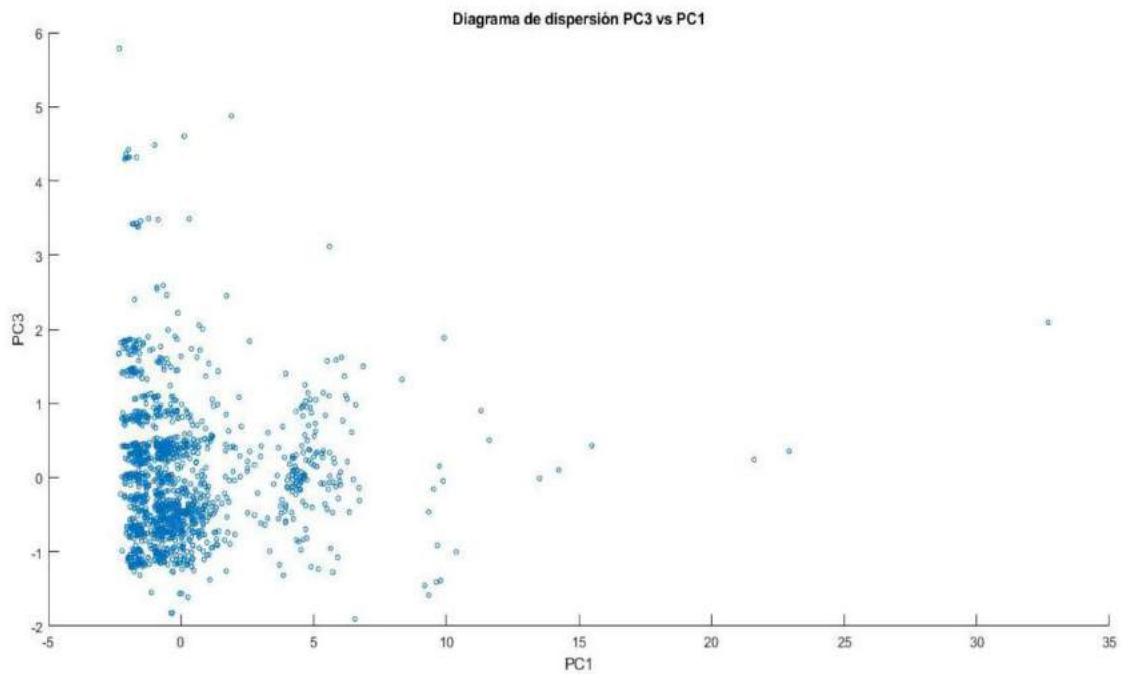
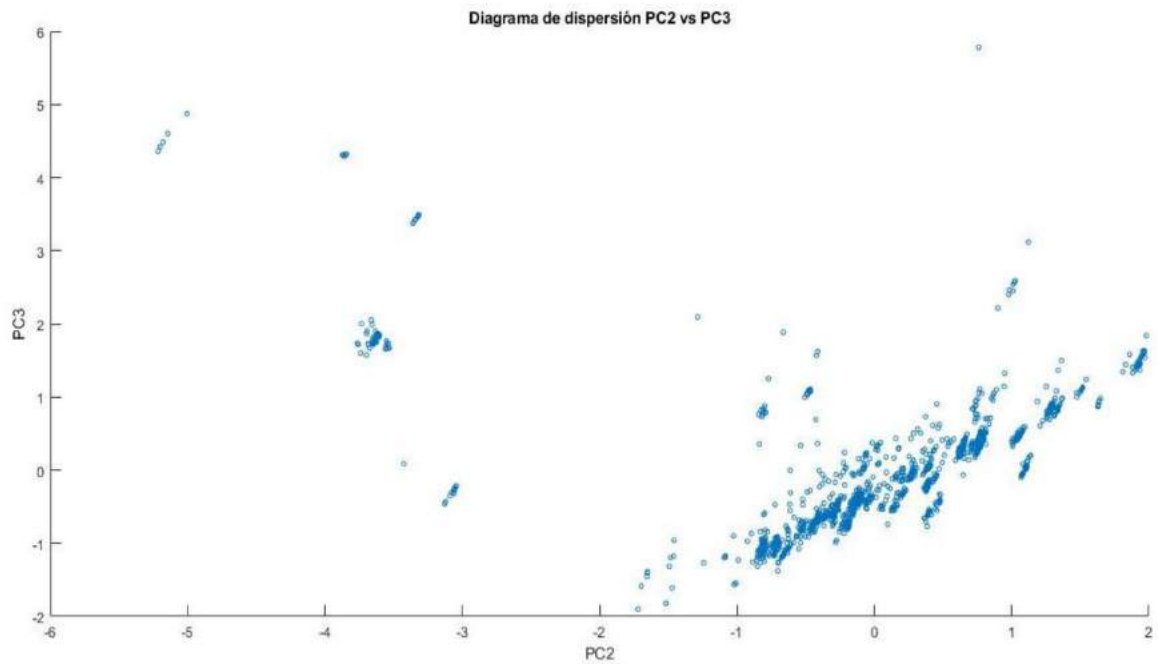


Figura 9: Diagrama de dispersión PC3 vs PC1 con todas las localidades.

De la misma manera con la figura 8, la figura 9 nos permite establecer el rango de valores en la componente 3. Para esta ultima componente las ZNI están dadas desde -2 a 6 aproximadamente.



*Figura 10: Diagrama de dispersión PC3 vs PC2 con todas las localidades.*

La figura 10, representa la otra vista posible del diagrama tridimensional, está vista no es tan útil puesto que está realizada a partir de la PC2 y PC3, que son las componentes que menos información almacenan, en comparación con la primera.

## 4. CLUSTERING

En la presente sección se detalla los criterios y pasos tenidos en cuenta para la conformación de las agrupaciones a partir de las características de cada una de las ZNI.

Como ya se mencionó anteriormente, los clústeres particionales son metodologías que requieren de procesos iterativos con el fin de minimizar o maximizar una función objetivo, es decir, son procedimientos más orientados a un modelo matemático definido. Entre estos clústeres particionales se encuentra un algoritmo que ha sido ampliamente utilizado, se denomina *K-Means*, la forma de proceder de este algoritmo es básicamente la que se menciona en la sección número 1. Sin embargo, de este algoritmo se han ido desarrollando múltiples variantes para contextos diferentes. Por ejemplo, si se retoma la definición del *centroide* definido en la expresión 1.3 se sabe que este representa el promedio de los individuos que conforman el clúster *k*. No obstante, esto no garantiza que el centroide sea un punto existente en los datos.

Por lo dicho anteriormente, se ha tomado en consideración para este estudio la implementación de la variante del *K-Means* que recibe el nombre de *K-Medoids* siendo el *medoide* un punto existente en la data, es decir, que el algoritmo de *K-Medoids* permite emplear una ZNI como representante del clúster, además debido a que este algoritmo es más resiliente a *outliers* comparado con el *K-Means* (Mirkin, 2005). El proceso de asignar el medoide tiene en cuenta la determinación del centroide, puesto que, a partir de conocer el centroide del clúster se calcula la distancia de todos los individuos al centroide y se elige como medoide el individuo más cercano al centroide.

### 4.1. PARAMETROS DEL CLUSTER

Para la clusterización en el estudio se empleó la función de Matlab *kmedoids* con los siguientes parámetros:

```
[idx,C,sumd]=kmedoids(XYZ,k,'Distance','ChooseDistance',"Algorithm","ChooseAlgorithm");
```

Donde:

*idx*= es un vector que almacena el número del clúster al que pertenece cada ZNI

*C*= es una matriz que almacena las características de los *k* medoides

*Sumd*= parámetro que permite observar cual fue el valor del SSE

*XYZ*= campo de entrada para ingresar la data de las 3 componentes principales

*K*= campo de entrada para ingresar el número de clústeres a realizar

Las observaciones de salida (*idx*, *C*, *sumd*) están claramente definidas, a diferencia de las de entrada. La lista de métricas se muestra más adelante, mientras que las consideraciones del tipo de algoritmo empleado para la función *kmedoids* se usa a partir del número de filas que tenga la matriz de entrada *XYZ*, entonces:

- Si el número de filas de *XYZ* es inferior a 3000, "pam" es el algoritmo por defecto.
- Si el número de filas está entre 3000 y 10000, 'small' es el algoritmo por defecto.



- Para todos los demás casos, "grande" es el algoritmo por defecto.

En este caso, *Partitioning Around Medoids (PAM)* es el algoritmo recomendado, es el algoritmo clásico para resolver la clusterización a partir de la metodología de kmedois, obteniendo individuos representativos (Kaufman & Rousseeuw, 2005). Después de aplicar la función de inicialización para seleccionar las posiciones iniciales de los medoides, el programa busca entre todos los posibles intercambios entre medoides y no medoides para ver si la suma de las distancias entre medoides y miembros del clúster disminuye.

El algoritmo procede como sigue.

- Paso de construcción: Cada uno de los k clústeres se asocia a un medoid potencial.
- Paso de intercambio: Dentro de cada clúster, cada punto se prueba como medoide potencial comprobando si la suma de las distancias dentro del clúster se reduce utilizando ese punto como medoide. Si es así, el punto se define como un nuevo medoide. A continuación, cada punto se asigna al clúster con el medoide más cercano.

El algoritmo repite los pasos de construcción e intercambio hasta que los medoides no cambien o se cumplan otros criterios de terminación.

#### 4.2.CRITERIOS DE SELECCIÓN DE LOS PARAMETROS DE ENTRADA A LA FUNCIÓN KMEDOIDS

Como se observó anteriormente, se debe ingresar el número de clústeres a realizar, la métrica de distancia y por último el tipo de algoritmo. A continuación se ahonda en cada una de estas inquietudes:

##### 4.2.1. Número óptimo de clústeres y selección de la métrica de distancia

El problema de estimar el número correcto de clústeres (k) es uno de los principales retos de la agrupación K-means y sus variantes. Varios investigadores han propuesto nuevos métodos para abordar este problema. A continuación se mostrarán cuatro de los más utilizados y disponibles para realizar en Matlab mediante una función definida que será detallada en las próximas líneas.

- **Calinski -Harabasz Index:** proponen definir el número óptimo de clústeres como el valor k que maximiza el estadístico que matemáticamente se expresa como:

$$CH(k) = \frac{\frac{B(k)}{k-1}}{\frac{W(k)}{N-k}} \quad (4.1)$$

Donde N representa el número de individuos, B(k) y W(k) son la suma de cuadrados entre e intra de los clústeres, respectivamente (con k clústeres).

- **Gap Statistic:** Este estadístico compara, para diferentes valores de k, la varianza total intra-cluster observada frente al valor esperado acorde a una distribución uniforme de referencia. La estimación del número óptimo de clústeres es el valor k con el que se

consigue maximizar el estadístico gap, es decir, encuentra el valor de k con el que se consigue una estructura de clústeres lo más alejada posible de una distribución uniforme aleatoria.

- **Davies Bouldin:** para este a diferencia del Calinski Index se propone obtener el k óptimo a partir de minimizar el estadístico.
- **Silhouette Coefficient:** este criterio se formula considerando las distancias entre e intra-clústeres. La expresión es la siguiente:

$$S = \frac{\sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)}}{N} \quad (4.2)$$

Donde la distancia de un individuo i al resto de los individuos que conforman su grupo es asignada a  $a_i$ , mientras que, la distancia de ese mismo individuo a los individuos que conforman los demás grupos se llama  $b_i$  y  $N$  representa el número de grupos.

Como ya se abordó, los parámetros anteriores se basan en el cálculo de la distancia entre individuos, por dicha razón es necesario evaluar que distancia entrega mejores resultados.

#### 4.3. FUNCIÓN *EVALCLUSTERS*

La selección del número óptimo de clústeres se enfoca en el cálculo de los indicadores teniendo en cuenta también la métrica de distancia empleada, para esta tarea se empleó la función de Matlab *evalclusters* que permite calcular los criterios con una sencilla instrucción, estructurada de la siguiente manera:

```
eva = evalclusters (XYZ, 'ClusterAlgorithm', 'ChooseCriterion', 'KList', [1:k])
```

Donde:

XYZ= Es la matriz que almacena los datos de las componentes principales.

k= Numero de clústeres a analizar

El algoritmo de clusterización a implementar en la función es el *K-Means*, la función no considera las variantes, sin embargo, la aproximación tiene mucha precisión. El resumen de los resultados se muestra a continuación:

Criterio	Valores de los criterios							Métrica de distancia
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	
<b>CH</b>	NaN	1855	1574	1741	<b>1889</b>	1736	1670	<i>Parámetro no valido para este indicador.</i>
<b>DB</b>	NaN	<b>0.484</b>	0.522	0.787	0.717	0.898	0.827	<i>Parámetro no valido para este indicador</i>
<b>GapS</b>	1.104	1.165	1.267	1.502	<b>1.667</b>	1.651	1.665	<i>SquareEuclidean</i>
<b>S</b>	NaN	<b>0.841</b>	0.816	0.564	0.567	0.552	0.487	
<b>GapS</b>	0.697	0.666	0.773	0.828	0.917	0.946	<b>0.965</b>	<i>Cityblock</i>
<b>S</b>	NaN	<b>0.586</b>	0.393	0.354	0.376	0.352	0.366	
<b>GapS</b>	-0.39	0.09	0.225	0.476	<b>0.532</b>	0.529	0.482	<i>Correlation</i>
<b>S</b>	NaN	0.828	0.871	<b>0.874</b>	0.843	0.771	0.751	

<b>GapS</b>	-0.4	-0.36	0.003	0.041	0.109	0.215	<b>0.261</b>	<i>Cosine</i>
<b>S</b>	NaN	0.626	0.674	<b>0.716</b>	0.657	0.653	0.659	

Tabla 5: Resultados de los parámetros usados para definir el valor de  $k$ .

Los parámetros resaltados en rojo son los mejores resultados para cada uno de los cálculos. El criterio de  $DB$  y  $S$  favorece la selección de dos clústeres, sin embargo, este número no garantiza una correcta diferenciación entre ellos, por lo cual es descartado del estudio. De manera similar, se descartan las opciones de  $k=6$  y  $k=7$  debido a que los criterios no las favorecen en lo absoluto, solo el criterio de  $S$  las considera como viables para los casos en donde se emplea la distancia *Cityblock* y *Cosine*. Por tal razón, se empleará también como método selector las vistas de las primeras dos componentes.

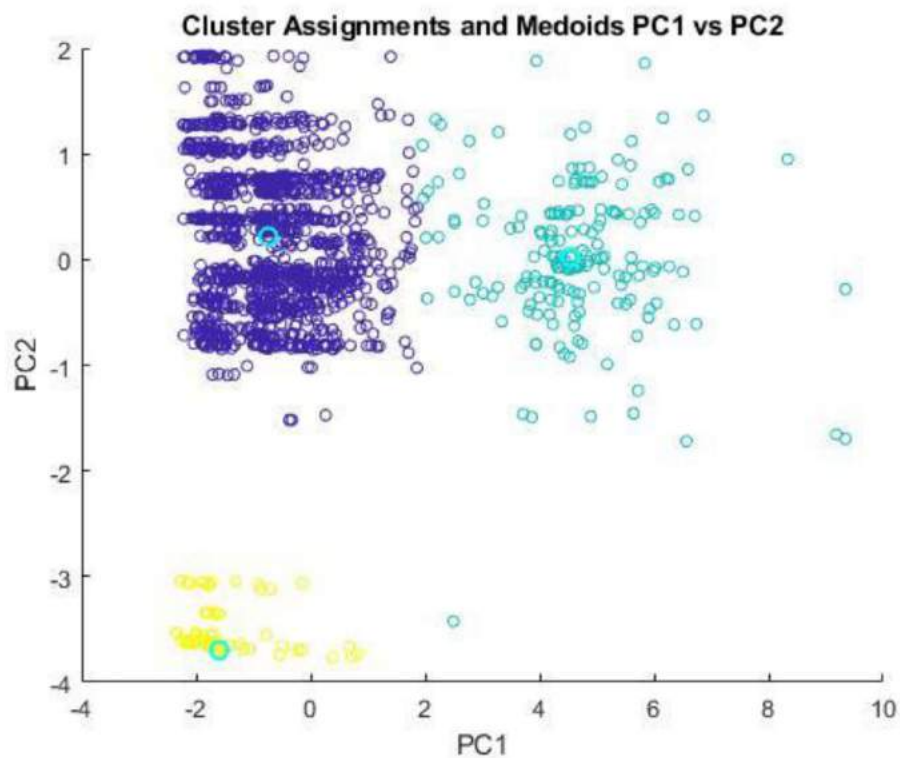


Figura 11: Gráfico de dispersión, vista  $PC2$  vs  $PC1$ , con la asignación de los grupos ( $k=3$ ).

La vista anterior refleja por qué los valores del criterio  $S$  son más elevados, puesto que, los individuos de los clústeres están más separados. Se puede observar que hay una clara diferenciación entre los 3 clústeres. Sin embargo, la distribución de zonas no es adecuada, el clúster azul oscuro cuenta con la mayor proporción de los datos, por tanto, se descarta esta opción.

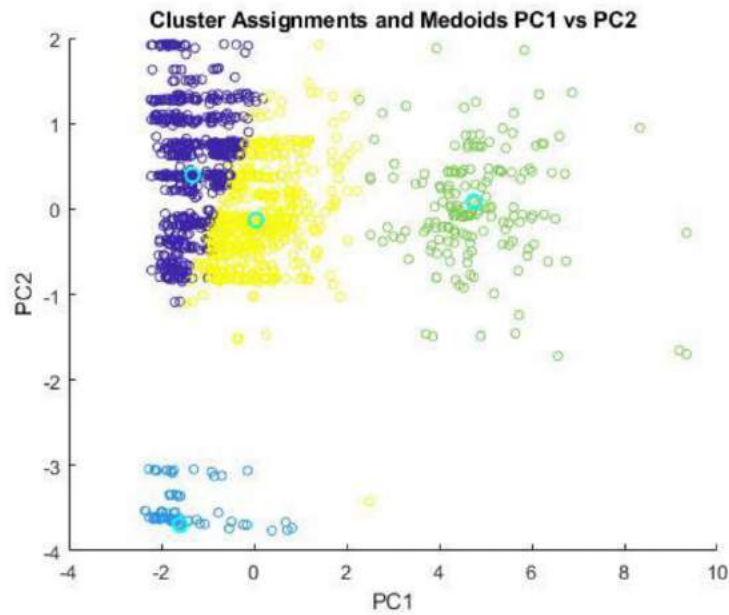


Figura 12: Gráfico de dispersión, vista PC2 vs PC1, con la asignación de los grupos ( $k=4$ ).

La vista con  $k=4$  representa una buena opción, a pesar de que los indicadores no la respaldan, solo el criterio  $S$  para las dos últimas métricas la defienden. Con esta distribución de los clústeres es posible diferenciarla sin embargo, el procedimiento es más complejo debido a que los límites entre el clúster azul oscuro y el amarillo se da a partir de una línea recta con pendiente positiva.

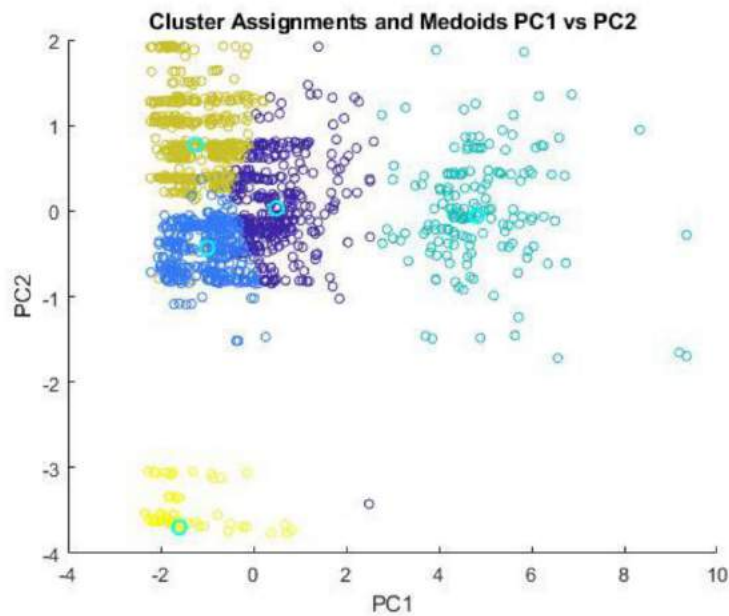


Figura 13: Gráfico de dispersión, vista PC2 vs PC1, con la asignación de los grupos ( $k=5$ ).

Esta última distribución ( $k=5$ ) está respaldada por el criterio de *CH* y el *GapS* para la distancia *Square Euclidean* y la conformación de los clústeres es más uniforme, teniendo en el C1 (azul oscuro) cuenta con 354 ZNI, el C2 (verde agua) con 166, el C3 (azul) con 433, el C4 (dorado) 522 y por último el C5 (amarillo) con 66. Es decir, representan un 22.97%, 10.77%, 28.1%, 33.87% y 4.28% respectivamente. Por las anteriores razones, esa es la configuración escogida para el estudio.

#### 4.4. PRESENCIA DE OUTLIERS

Un *Outlier* o valor atípico es una observación que es numéricamente distante del resto de los datos. Al momento de clusterizar con la configuración escogida se obtuvo:

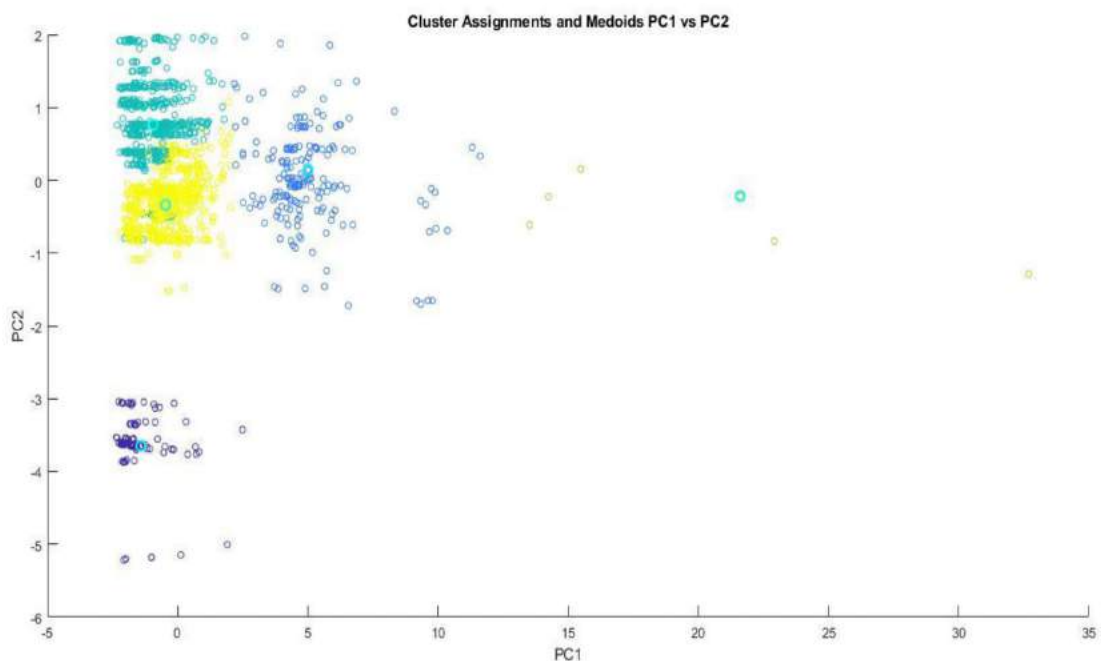


Figura 14: Gráfico de dispersión, grupos ( $k=5$ ), de los datos con outliers, vista PC2 vs PC1.

Con los *Outliers* se obtiene un clúster formado con solo seis ZNI, gráficamente se logra observar claramente que estas ZNI están representando un dato atípico. Para tratar los datos atípicos se han definido múltiples metodologías que tienen en cuenta diversos estadísticos como el promedio, la mediana y los cuartiles. La metodología empleada usa los cuartiles para identificar dichos outliers debido a que se ha demostrado que este método es útil cuando la data no está normalmente distribuida, como es este caso:

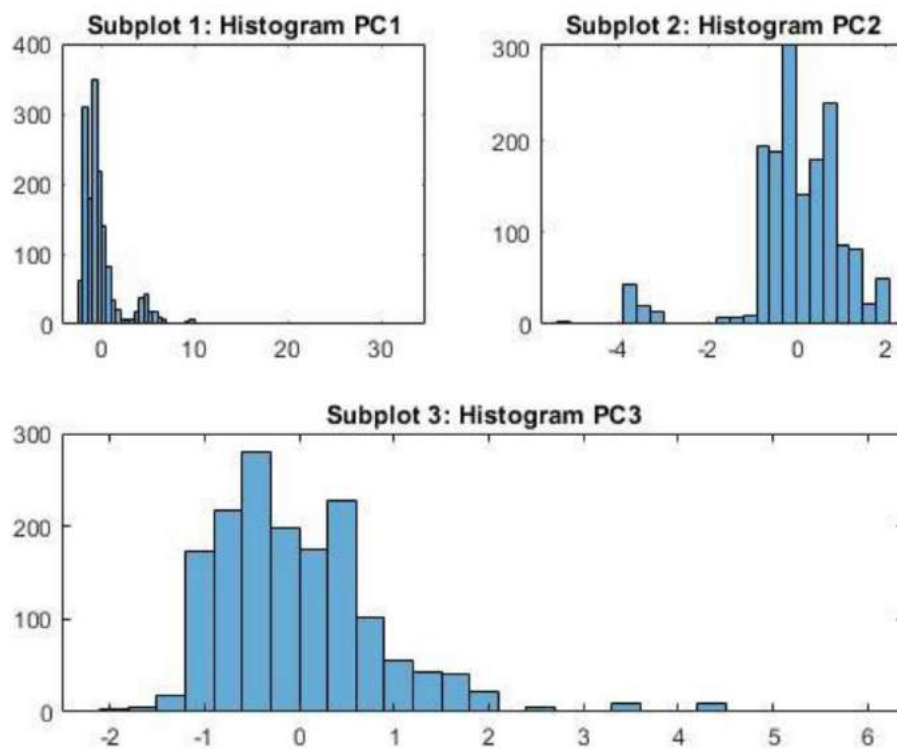
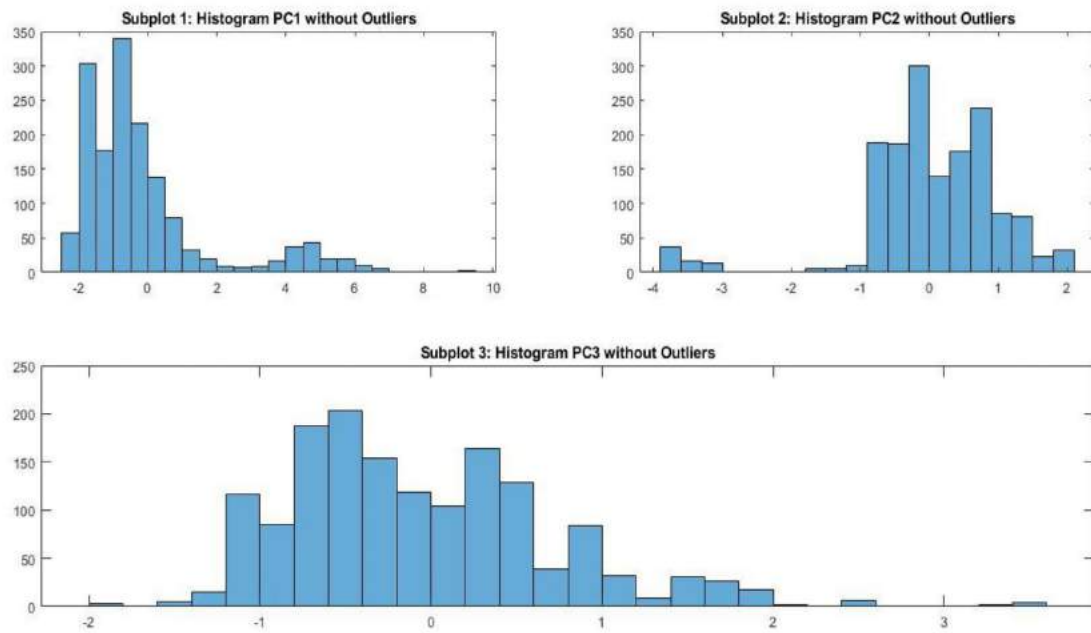


Figura 15: Histogramas para cada una de las componentes principales seleccionadas, presencia de outliers. Este procedimiento fue realizado a partir de la función de Matlab *rmoutliers* con los siguientes argumentos:

```
[scorescodWO, TF] = rmoutliers(scorescod(:,1:3), 'percentiles', [0 99]);
```

Con esa instrucción se obtiene la nueva data sin *outliers* y un vector lógico *TF* que permite conocer cuales individuos fueron eliminados. Para lo anterior, fue necesario ingresar la matriz (*scorescod*) con los valores correspondientes a las tres primeras componentes y el criterio de eliminación fue las ZNI que superen el percentil 99. Con eso se removieron 48 individuos, quedando en total 1540 con la siguiente distribución de frecuencias:



*Figura 16: Histogramas para cada una de las componentes principales seleccionadas, sin outliers.*  
 De aquí se entiende porque las vistas resultantes del PCA muestran una escala diferente a las del clúster, debido a que las gráficas mostradas anteriormente de las agrupaciones ya muestran los resultados sin *outliers* y cuentan con las escalas mostradas en la gráfica anterior, PC1 desde -2 a 10, PC2 desde -4 a 2 y finalmente PC3 desde -2 a 4.

## 5. POTENCIAL ENERGÉTICO DE BIOMASA

La determinación del potencial energético de la biomasa parte de la base de datos encontrada, sin embargo, como esta cuenta con datos desde 2007 hasta 2018, es necesario implementar una herramienta de pronóstico para poder estimar los valores referentes a los cultivos para el periodo de análisis, septiembre 2020. Como ya se mencionó anteriormente, la data original debe reestructurarse para que posea valor en el estudio, las consideraciones son las siguientes:

- Se extraen solo los datos para los municipios que cuentan con localidades en la base de datos para septiembre 2020.
- Se extraen solo los datos de los cultivos que son considerados como cultivos energéticos y están definidos en el atlas de biomasa de Colombia. Como lo muestra la siguiente figura:

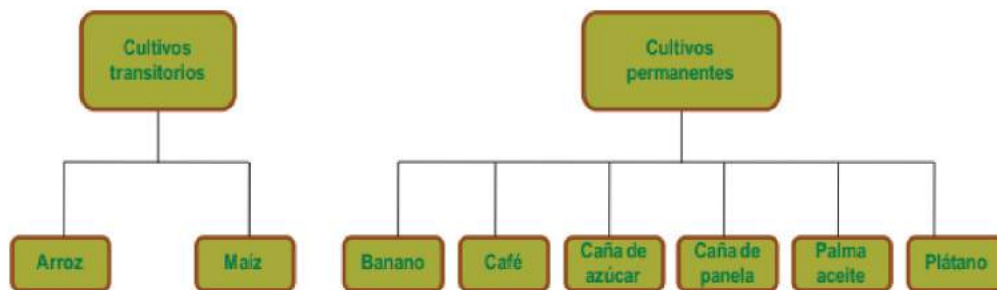


Figura 17: Cultivos con potencial energético considerable en Colombia (UPME, IDEAM, COLCIENCIAS, UIS).

- Se toman los valores desde el año 2016 a 2018 con el fin de tener 3 periodos de tiempo y que sean base para calcular los pronósticos para el 2020.

Los cultivos producidos en los 52 municipios que cuentan con localidades para el mes de septiembre de 2020 están distribuidos en arroz, maíz, bananito, banano, plátano, caña panelera y palma de aceite, ninguno de los municipios cuenta con hectáreas para la producción de caña de azúcar o café. De estos, la base de datos reporta, el área sembrada, la cosechada, la producción y el rendimiento. Sabiendo que el potencial energético será calculado a partir de la metodología que propone el atlas de biomasa:

$$PE (kWh) = M_{rs} * PCI/3600 \quad (5.1)$$

Donde:

$M_{rs}$ : es la masa de residuo seco en kg

$PCI$ : es el poder calorífico inferior kJ/kg

El factor de 1/3600 se usa para pasar de kJ a kWh, magnitud equivalente a la de la energía facturada para las zonas. De la expresión 5.1 se requiere conocer los PCI para cada uno de los cultivos así como también la cantidad de residuos. Para esto primero, se tienen los siguientes valores:



CULTIVO	PCI [kJ/kg]	OBSERVACIONES	REFERENCIA
<b>Maíz</b>	12799.55	Se obtiene de termoflex a partir de una composición de biomasa dada en la referencia, que está en base seca: <i>13788 kJ/kg</i> . Posteriormente se corrigió con una humedad de 6.98% dando como resultado el valor reportado.	Corn Stover PROXIMATE ANALYSIS WB (%) Moisture 5 Ash 6,98 Volatile 74,2 Fixed Carbon 13,8  (Lu, y otros, 2006)
<b>Plátano, banano y bananito</b>	2519.18	Se toma como referencia el PCI para el plátano seco, <i>2960 kJ/kg</i> obtenido de la referencia y corregido a partir del análisis próximo.	Poder calorífico obtenido de las muestras realizadas dentro del laboratorio de físico química de la Universidad de Cuenca en Ecuador. (Romero, 2019)
<b>Arroz</b>	13162.1	Calculado a partir de la ponderación del PCI para la paja obtenido de la referencia ( <i>14570 kJ/kg</i> ), el mostrado en la referente es PCS por tanto, se corrigió a partir de la entalpia de vaporización, la humedad y el contenido de H <sub>2</sub> del análisis último ( <i>13049.65 kJ/kg</i> ) y para la cascarilla ( <i>14483.65 kJ/kg</i> ) se tomó del Atlas de Biomasa de Colombia.	Para la paja de arroz. (Munder, Karaj, Gummert, Haefele, & Muller, 2012)  Para la cascarilla. (UPME, IDEAM, COLCIENCIAS, UIS)
<b>Palma de aceite</b>	12789.8	Obtenido del Atlas de Biomasa Residual de Colombia y fue ponderado a partir del potencial energético de la palma de aceite y las toneladas de residuo anuales.	Palma de aceite (UPME, IDEAM, COLCIENCIAS, UIS)
<b>Caña panelera</b>	11700	Reportado en la referencia. Manual para cálculo y diseño de un trapiche panelero.	(Bello Vasquez & Caicedo Medina, 1996)

Tabla 6: Valores del PCI empleados para el cálculo del potencial energético de cada cultivo.

Con lo anterior se reporta los valores de PCI utilizados en el desarrollo de los potenciales energéticos de biomasa. Para completar y poder implementar la expresión 5.1 es necesario calcular la masa de residuos. Como se tiene la producción de producto, reportada en la base de datos, entonces, se emplean los factores de residuo dados por el *Atlas de Biomasa Residual de Colombia*. Los valores empleados se muestran a continuación:

Cultivo	Tipo de residuo	Factor de residuo [toneladas de residuo/toneladas de producto]	Factor de residuo empleado
Maíz	Rastrojo	0.93	1.41
	Tusa	0.27	
	Capacho	0.21	
Plátano, banano y bananito	Raquis de banano	1	6
	Vástago de banano	5	
Arroz	Tamo	2.35	3
	Cascarilla	0.65	
Palma de aceite	Cuesco	0.22	0.85
	Fibra	0.63	
Caña panelera	Bagazo	2.53	6.28
	Hojas – Cogollo	3.75	

Tabla 7: Factores de residuo empleados para calcular la generación de residuos de cada cultivo.

Los residuos son calculados entonces a partir de la multiplicación entre las toneladas de producto dadas en la base de datos y el factor de residuo mostrado en la tabla anterior. De esta manera, es posible calcular el potencial energético para cada uno de los municipios de cada cultivo y finalmente la suma de estos, representan el potencial energético total del municipio. Cabe mencionar que para los cultivos transitorios (arroz y maíz) el área empleada para sembrar los cultivos fue un promedio de los valores en el año (2 periodos). A partir de lo anterior se plantean los pronósticos.

### 5.1. PRONOSTICO PARA 2020

Pronosticar el área sembrada y la cantidad de residuos para el año en cuestión, es de vital importancia pues estos parámetros serán empleados para el cálculo del potencial energético en cada municipio y para la asignación del potencial disponible para cada localidad. Para lograr dicho objetivo se empleó la función de Excel que recibe el nombre de *Pronostico.ETS*, esta calcula o predice un valor futuro en base a valores (históricos) existentes mediante el algoritmo de Suavizado exponencial triple (ETS). El valor pronosticado es una continuación de los valores históricos de la fecha de destino especificada, que debería ser la continuación de la línea de tiempo (Microsoft, 2021).

Esta función requiere que la línea de tiempo esté organizada con un paso constante entre los diferentes puntos. Por ejemplo, podría ser una línea de tiempo con valores mensuales los días 1 de cada mes, una línea de tiempo anual o una línea de tiempo de índices numéricos. La sintaxis de la función es la siguiente:

*PRONOSTICO.ETS(fecha – destino , valores, línea de tiempo)*

Siendo,

Fecha-destino: argumento obligatorio. Es el punto de datos cuyo valor se desea predecir. La fecha de destino puede ser fecha, hora o numérico. Si la fecha de destino es cronológicamente antes del final de la línea de tiempo histórica, PRONOSTICO.ETS devuelve el error #NUM!.

Valores: argumento obligatorio. Los valores son los valores históricos para los que desea pronosticar los puntos siguientes.

Escala de tiempo: argumento obligatorio. Es la matriz o el rango de datos numéricos independientes. Las fechas de la escala de tiempo deben tener un paso coherente entre ellas y no pueden ser cero. No es necesario ordenar la línea de tiempo, ya que PRONOSTICO.ETS la ordenará de forma implícita para los cálculos. Si no se puede identificar un paso constante en la escala de tiempo proporcionada, PRONOSTICO.ETS devolverá el error #NUM! . Si la escala de tiempo contiene valores duplicados, PRONOSTICO.ETS devolverá el error #¡VALOR! . Si los rangos de la escala de tiempo y los valores no son del mismo tamaño, PRONOSTICO.ETS devolverá el error #N/A.

La metodología empleada para el pronóstico del conjunto de datos es predecir el valor del área sembrada y el rendimiento del cultivo, a partir de estos dos se calcula la producción de toneladas de producto, mediante la multiplicación de estos dos. Los residuos se obtienen de la misma manera anteriormente descrita, empleando los factores de residuos mostrados en la tabla 7 y el potencial usando los PCI reportados en la tabla 6. Para ser más precisos y coherentes con los resultados de la función para predecir el rendimiento se establecieron una serie de condicionales:

- El resultado NO puede ser negativo, este problema se presentó debido a que en algunos cultivos la tendencia iba decreciendo. Así mismo, se observó que en algunos casos el pronóstico era útil teniendo en cuenta el valor absoluto del resultado obtenido.
- Se definieron límites del valor de rendimiento a partir de referencias, por consiguiente, se estableció un valor máximo del rendimiento de cada cultivo, tabla 8.

CULTIVO	RENDIMIENTO [ton prod/ha semb]	OBSERVACIONES	REFERENCIA
Maíz	3.6	Rendimiento promedio nacional. En Colombia, el maíz es el tercer cultivo con mayor superficie de siembra después del café y el arroz.	Maíz para Colombia visión 2030. (Centro Internacional de Agricultura Tropical (CIAT), Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT, 2019)
Plátano, banano y bananito	23.98	Rendimiento promedio del 2014, histórico desde el 2007. Tendencia	(Ministerio de Agricultura, 2014)

		levemente decreciente.	
<b>Arroz</b>	7.5	Rendimiento promedio nacional esperado por Fedearroz. Corresponde también por el rendimiento obtenido en Huila para el 2017.	(Grupo Bancolombia, 2018)
<b>Palma de aceite</b>	No se consiguieron registros.		
<b>Caña panelera</b>	14	Rendimiento máximo del 2014, histórico desde el 2007. Tendencia creciente comparado con los valores reportados en la base de datos para los años desde 2016 a 2018.	(Ministerio de Agricultura, 2014)

*Tabla 8: Valores máximos de rendimiento utilizados para el pronóstico del año 2019 y 2020.*

Con lo anteriormente definido, fue posible completar los datos para el año 2020.

## **5.2. ASIGNACIÓN DE POTENCIAL ENERGETICO BRUTO PARA CADA ZONA**

Los datos disponibles hasta el momento son cantidades del municipio y para poder implementar esto al estudio es necesario tener dichos potenciales para cada una de las zonas, pues todas han sido tratadas como individuos. Por lo cual se realizó la asignación en base a la cantidad de usuarios, para esto se calculó el total de los usuarios para cada municipio y con este se encuentra la proporción de usuarios que aporta cada localidad. De esta manera, al multiplicar el potencial del municipio con la proporción de usuarios que aporta cada municipio, se logran obtener las estimaciones del potencial energético anual de cada zona tanto total como para cada cultivo. Este tipo de ponderación considera que la biomasa es accesible para cada una de las zonas y usuarios, sin tener en cuenta la ubicación del cultivo.

## **5.3. POTENCIAL TECNICO Y MARGENES DE BIOMASA**

En los estados del arte para la gasificación de biomasa a pequeña escala, los sistemas de generación con gasificador del sistema downdraft y motor de combustión interna tienen eficiencias comprendidas entre el 12.5% y 23%. Siendo el valor más pequeño obtenido a partir de sistemas de gasificación que emplean biomasa producida a partir de cascarilla de arroz, mientras que los valores intermedios y más altos se obtuvieron usando madera como combustible (Bocci, y otros, 2014). Por lo cual, para encontrar el potencial técnico se considerará una eficiencia de conversión de energía del 13%. Entonces, el potencial técnico de biomasa para cada localidad se expresa como:

$$PB_{técnico} = PB_{brutoZNI} * \varepsilon_{sistema_{gasificación}} = 0.13 * PB_{brutoZNI} \quad (5.2)$$

Donde:

$PB_{técnico}$ : indica el potencial de biomasa técnico.

$PB_{brutoZNI}$ : potencial de biomasa bruta para cada ZNI.

$\varepsilon_{sistema_{gasificación}}$ : eficiencia de conversión de un sistema de gasificación downdraft.

Por otra parte, se calcula la proporción de energía que es posible generar a partir del potencial técnico de la biomasa en relación con la energía de diseño (energía con la cual se supe el consumo de subsistencia para todos los usuarios) para cada localidad. Esta proporción recibe el nombre de margen de biomasa y se expresa de la siguiente manera:

$$BM_{ZNI} = \frac{PB_{técnicoZNI}}{E_{diseñoZNI}} \quad (5.3)$$

Donde:

$BM_{ZNI}$ : margen de biomasa para cada ZNI.

$E_{diseñoZNI}$ : energía de diseño para cada ZNI, expresa la cantidad de energía para suplir el consumo de subsistencia de cada uno de los usuarios.

Si el valor obtenido es menor a 1 representa que a partir de la biomasa es posible solo generar una fracción de la energía total, si es igual a 1 es posible generar toda la energía requerida de la zona a partir de la biomasa y si es mayor es posible generar más del consumo de subsistencia.

El cálculo de los márgenes se discreta según si la biomasa es seca o húmeda. Considerando como biomasa seca las obtenidas a partir del arroz, maíz, caña y palma. Por otro lado, el margen de biomasa húmeda calculado con los potenciales técnicos de la biomasa del banano y plátano. Entonces:

$$DBM_{ZNI} = \frac{(PB_{técZNI_{Arroz}} + PB_{técZNI_{Maíz}} + PB_{técZNI_{Palma}}) * FR_{agr} + PB_{técZNI_{Caña}} * FR_{caña}}{E_{diseñoZNI}} \quad (5.4)$$

Donde:

$DBM_{ZNI}$ : es el margen de biomasa dado por la biomasa seca, compuesto por los residuos provenientes de los cultivos de arroz, maíz, palma de aceite y la caña panelera.

$FR_{agr}$ : representa el factor de recolección de residuos para los cultivos estudiados, a excepción de la caña panelera.

$FR_{caña}$ : factor de recolección de residuos empleado para la caña panelera.

$$WBM_{ZNI} = \frac{(PB_{técnicoZNIPlátano} + PB_{técnicoZNIBanano}) * FR_{agri}}{E_{diseñoZNI}} \quad (5.5)$$

$WBM_{ZNI}$ : margen de biomasa húmeda, compuesta por la biomasa del plátano y banano.

El término FR mostrado en las expresiones 5.4 y 5.5 se refiere al factor de recolección de los residuos, se incluye con el fin de mostrar de manera más real los márgenes de biomasa, puesto que, los residuos generados en los cultivos no son recolectados al 100%. En literatura, es común encontrar valores de 0.8, es decir, que el 80% de los residuos generados son recolectados para la biomasa agrícola (Karaj, Rehl, Leis, & Muller, 2010), este valor coincide con reportes para el plátano en los cuales están entre el 70 y el 80% (Pérez Saldarriaga & Salazar Briceño, 2015). Mientras que para el caso de la biomasa proveniente de la caña panelera se reporta entre un 4 y 5% (Bello Vasquez & Caicedo Medina, 1996). Como tal, los valores usados fue un 5% para la caña panelera y un 80% para los demás cultivos.

El margen de biomasa total se define como la suma del margen de biomasa seca y húmeda.

$$TBM_{ZNI} = DBM_{ZNI} + WBM_{ZNI} \quad (5.6)$$

Donde:

$TBM_{ZNI}$ : es el margen de biomasa total, resultante de la suma de los márgenes de biomasa seca y húmeda.

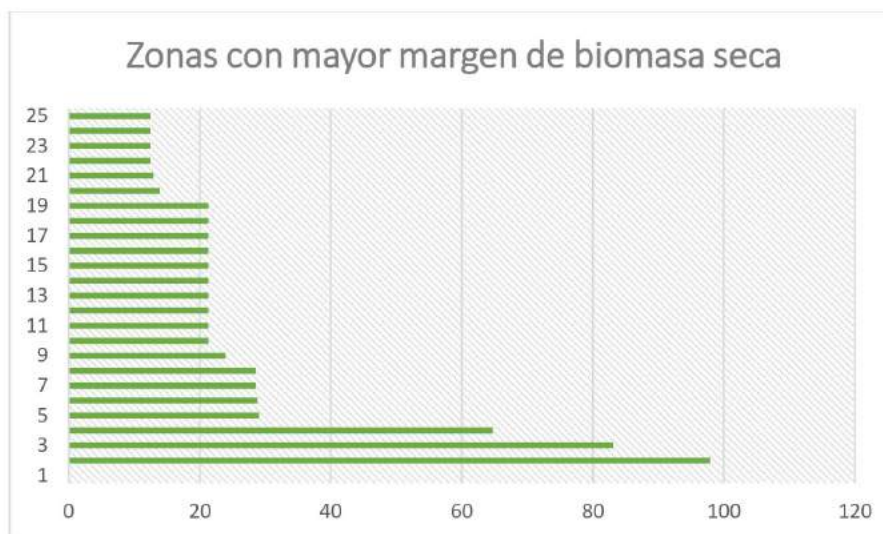


Figura 18: Zonas con mayor margen de biomasa seca (DBM), 25 primeras.

De los resultados obtenidos, se muestra por ejemplo las 25 zonas con mayores márgenes de biomasa seca. La figura 18 muestra las 25 zonas con mayor margen de biomasa seca (zonas eje Y, valor del DBM eje X) de las cuales, 17 de ellas pertenecen al municipio de RIOSUCIO

y la composición del margen está dado por los siguientes aportes de cada cultivo, como se muestra a continuación:

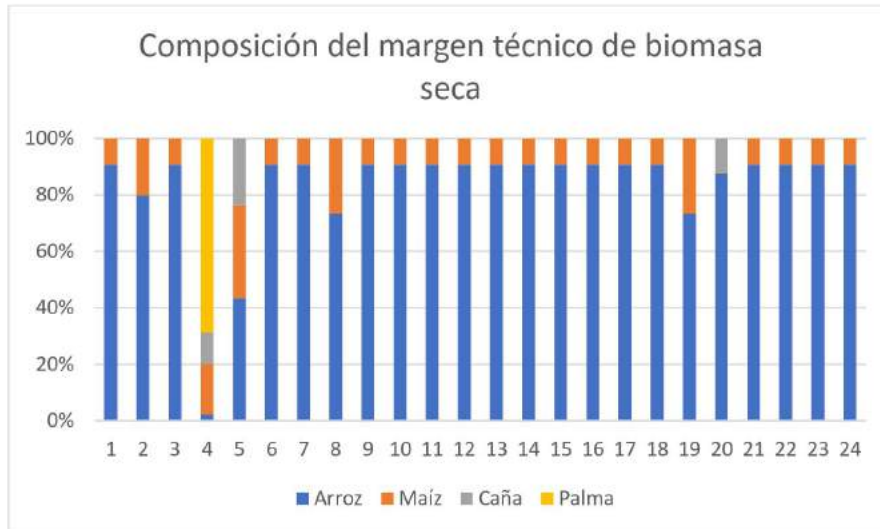


Figura 19: Composición del margen de biomasa seca (DBM) para las 25 zonas con mayor valor, aporte de cada cultivo.

Como se puede observar, en la mayoría de los casos, el margen de biomasa seca está compuesto mayormente de la energía que es posible generar a partir del arroz. De las 25 solo hay una en la cual el margen de biomasa seca está dado por otro cultivo diferente, que es el de palma de aceite, esta zona está ubicada en el departamento de Vichada, municipio La Primavera, Santa Barbara, cuenta con 6000 hectáreas sembradas de palma de aceite.

## 6. ANALISIS DE RESULTADOS

En esta sección, se trabajan los datos y resultados obtenidos en cada una de las secciones anteriores. Todos con el fin de poder definir nuevas tipologías de las ZNI como ya se mencionó en algún momento, dejando atrás las actuales dadas solo por el número de usuarios.

### 6.1. TIPOLOGIAS DE LAS CARACTERISTICAS DE LAS ZNI

La clusterización por medio de Matlab da como resultado la conformación de 5 grupos como ya vimos en la sección número cuatro. Esta solo indica por cuales zonas están compuestos los conglomerados, no obstante, no entrega información referente a las características que posee cada clúster, con el fin de conocerlas se implementó el análisis de frecuencias para datos agrupados.

Para poder comparar los resultados obtenidos para cada clúster, se presenta los valores máximos y mínimos de las variables de estudio (implementadas en el PCA) de todas las zonas, es decir, representan los mínimos y máximos globales del conjunto de datos.

VALORES REALES		
VARIABLES	MINIMO GLOBAL	MAXIMO GLOBAL
<b>CAP</b>	3	350
<b>USU</b>	8	381
<b>EFACT</b>	240.24	19637.37
<b>CU</b>	978.27	2023.82
<b>FACT</b>	287600	29904426.8
<b>SUBS</b>	230640	25386382.6
<b>DIE</b>	26	2197
<b>PDIE</b>	6312.49	12000
<b>PTDIE</b>	2000	3101699

*Tabla 9: Mínimos y máximos globales de los valores reales para las variables que caracterizan la prestación de energía.*  
El análisis de frecuencia para datos agrupados se estableció con un número de intervalos de 10, iniciando desde el valor mínimo de cada variable para cada clúster y finalizando en el máximo. Por tanto, la amplitud de los intervalos se expresa:

$$A = \frac{V_{max}(X) - V_{min}(X)}{n} \quad (6.1)$$

Donde:

$A$ : es la amplitud del intervalo.

$V_{max}(X)$ : es el valor máximo de la variable  $X$ .

$V_{min}(X)$ : es el valor mínimo de la variable  $X$ .

$n$ : es el número de intervalos definido.



La estructura de la tabla de frecuencias es la típica, a continuación se muestra la realizada para la variable capacidad de las zonas pertenecientes al clúster 1 (C1):

INTERVALO	INFERIOR	SUPERIOR	fi	fr%	Fr acum &
1	20	38	29	8.192%	8.192%
2	38	56	103	29.096%	37.288%
3	56	74	94	26.554%	63.842%
4	74	92	71	20.056%	83.898%
5	92	110	24	6.780%	90.678%
6	110	128	12	3.390%	94.068%
7	128	146	7	1.977%	96.045%
8	146	164	7	1.977%	98.023%
9	164	182	6	1.695%	99.718%
10	182	200	1	0.282%	100.000%
			354	1	

Tabla 10: Tabla de frecuencias para la variable CAPACIDAD de las zonas que conforman el C1.

La información anterior, también se muestra a partir de:

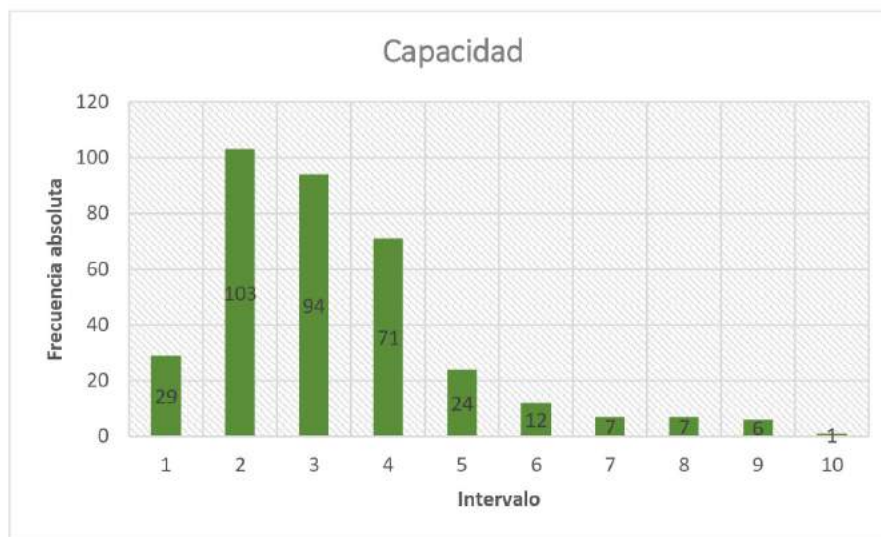


Figura 20: Histograma para cada uno de los intervalos definidos para la variable CAPACIDAD de las zonas que conforman el C1.

Como se observa una gran mayoría de las zonas que conforman el clúster se concentran en algunos intervalos específicos. Por tanto, para definir las tipologías fue necesario sesgar los intervalos de las variables a partir de establecer un rango característico. Este rango se conforma por los intervalos que sumados entre sí aportan una frecuencia relativa acumulada mayor o igual al 70%. Adicionalmente, es necesario que estos intervalos sean continuos. Para el caso mostrado en la tabla y figura 16 el rango característico está dado por los intervalos del 1 al 4, en los cuales se logra abarcar un 83.89% de las zonas que conforman el C1 (354).

A continuación se muestran las tipologías encontradas para cada uno de los conglomerados.

VALORES REALES – C1				
VARIABLES	RANGO CARACTERISTICO		%Zonas	Descripción
<b>CAP</b>	20	92	83.90%	Capacidad media baja
<b>USU</b>	56	122.0	86.16%	Usuarios medios bajos
<b>EFACT</b>	2280.0	5515.2	92.09%	Energía media baja
<b>CU</b>	1291.9	1710.2	88.42%	CU medio
<b>FACT</b>	4141026.3	7710864.1	84.46%	Fact medio baja
<b>SUBS</b>	3390545.5	6345321.6	83.05%	Fact medio baja
<b>DIE</b>	245.0	548.9	90.11%	Diesel medio
<b>PDIE</b>	6312.5	6950.0	96.05%	Bajo
<b>PTDIE</b>	348908.2	692269.4	81.07%	Bajo

Tabla 11: Tipología de las zonas que conforman el clúster 1 (C1).

El clúster 1 se encuentra conformado con 354 ZNI que representan un 22.97% del total, los rangos característicos permiten definir con mucha precisión las características de generación en dichas localidades. A su vez esto también se demuestra si se observan las características del medoide del clúster:

Medoide C1: Caleta viento libre, San Andrés del Tumaco, Nariño.								
CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
45	102	4593	1440.2	6596061	5434108	497	6724	474930

Tabla 12: Valores reales de la ZNI representativa, medoide del C1.

VALORES REALES – C2				
VARIABLES	RANGO CARACTERISTICO		%Zonas	Descripción
<b>CAP</b>	30	158	86.14%	Capacidad media
<b>USU</b>	151	220.0	86.14%	Usuarios medios
<b>EFACT</b>	9126.9	14382.2	81.33%	Energía medio alta
<b>CU</b>	1349.4	1734.8	77.11%	CU medio
<b>FACT</b>	11407239.8	21977061.0	86.75%	Fact media alta
<b>SUBS</b>	9675914.5	18653324.9	84.94%	Fact media alta
<b>DIE</b>	943.4	1413.5	84.34%	Diesel alto
<b>PDIE</b>	6312.5	6878.7	88.55%	Bajo
<b>PTDIE</b>	1249010.0	2175354.5	81.33%	Alto

Tabla 13: Tipología de las zonas que conforman el clúster 2 (C2).

El clúster 2 se encuentra conformado por 166 ZNI que representan un 10.77% del total, los rangos característicos permiten definir con mucha precisión las características de generación en dichas localidades. A su vez esto también se demuestra si se observan las características del medoide del clúster:

Medoide C2: La loma, Santa Barbara, Nariño.								
CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
65	167	12692	1410	17905239	14694543	1249	6799	1710031

Tabla 14: Valores reales de la ZNI representativa, medoide del C2.

VALORES REALES – C3				
VARIABLES	RANGO CARACTERISTICO		%Zonas	Descripción
<b>CAP</b>	3	69.8	90.07%	Poca capacidad
<b>USU</b>	27.6	66.6	84.30%	Pocos usuarios
<b>EFACT</b>	726.3	3311.4	89.15%	Poca energía fact
<b>CU</b>	1203.2	1441.3	90.07%	Cu medio bajo
<b>FACT</b>	886736.7	4019325.6	84.30%	Fact baja
<b>SUBS</b>	699446.3	3172289.0	81.52%	Fact baja
<b>DIE</b>	38.0	349.0	95.15%	Medio bajo
<b>PDIE</b>	6501.2	6878.7	93.53%	Bajo
<b>PTDIE</b>	56913.3	432206.4	86.84%	Medio bajo

Tabla 15: Tipología de las zonas que conforman el clúster 3 (C3).

El clúster 3 se encuentra conformado por 433 ZNI que representan un 28.1% del total, los rangos característicos permiten definir con mucha precisión las características de generación en dichas localidades. A su vez esto también se demuestra si se observan las características del medoide del clúster:

Medoide C3: Mesopotamia, Bojayá, Choco								
CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
36	52	2340	1343.8	3144492	2509884	252	6702.11	86254.43

Tabla 16: Valores reales de la ZNI representativa, medoide del C3.

VALORES REALES – C4				
VARIABLES	RANGO CARACTERISTICO		%Zonas	Descripción
<b>CAP</b>	4	47.8	85.63%	Poca capacidad
<b>USU</b>	20	70.0	88.70%	Pocos usuarios
<b>EFACT</b>	563.2	2820.2	87.16%	Poca energía
<b>CU</b>	1480.4	1860.8	88.12%	CU medio alto
<b>FACT</b>	996533.7	4540573.2	82.18%	Fact medio baja
<b>SUBS</b>	844562.4	3868564.4	81.99%	Fact medio baja
<b>DIE</b>	75.1	320.6	89.46%	Medio bajo
<b>PDIE</b>	6501.2	6878.7	88.51%	Bajo
<b>PTDIE</b>	85816.4	423906.8	87.93%	Medio bajo

Tabla 17: Tipología de las zonas que conforman el clúster 4 (C4).

El clúster 4 se encuentra conformado por 522 ZNI que representan un 33.87% del total, los rangos característicos permiten definir con mucha precisión las características de generación en dichas localidades. A su vez esto también se demuestra si se observan las características del medoide del clúster:

Medoide C4: Uyaba, Medio Baudó, Choco								
CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
25	48	1584	1737.66	2752453.44	2322888.48	307	6668	196667

Tabla 18: Valores reales de la ZNI representativa, medoide del C4.

VALORES REALES – C5				
VARIABLES	RANGO CARACTERISTICO		%Zonas	Descripción
CAP	5	43	75.76%	Poca capacidad
USU	8	56.0	87.88%	Pocos usuarios
EFACT	240.2	1704.1	77.27%	Poca energía
CU	978.3	1233.5	90.91%	CU bajo
FACT	287600.0	2478164.5	83.33%	Fact Baja
SUBS	230640.0	1987320.0	87.88%	Fact Baja
DIE	26.0	197.9	80.30%	Bajo
PDIE	9000.0	11100.0	90.91%	Alto
PTDIE	2000.0	816028.5	86.36%	Medio

Tabla 19: Tipología de las zonas que conforman el clúster 5 (C5).

El clúster 5 se encuentra conformado por 66 ZNI que representan un 4.28% del total, los rangos característicos permiten definir con mucha precisión las características de generación en dichas localidades. A su vez esto también se demuestra si se observan las características del medoide del clúster:

Medoide C5: Santa Rita, Inírida, Guainía								
CAP	USU	EFACT	CU	FACT	SUBS	DIE	PDIE	PTDIE
80	18	604.8	1197.1	724012.12	580608	70.6764984	10920	315902.745

Tabla 20: Valores reales de la ZNI representativa, medoide del C5.

Este clúster 5 al ser el que menos localidades lo conforman, es normal que el medoide no represente con total exactitud los rangos característicos, como se observa con la variable de capacidad instalada.

De los rangos mostrados anteriormente, es notable que el C1 y C2 tienen un comportamiento único y claramente diferenciable, sin embargo, el C3, C4 y C5 presenta similitud en sus características. Se puede decir que C5 se diferencia pues su rango de generación es más limitado a pesar de que cuenta con un CU bajo, sin embargo, los costos del diésel son elevados. Por otro lado, C3 cuenta con menor variación en el CU, junto con un menor costo, lo que le permite tener un rango más alto de Efact. Para tratar de caracterizar de mejor manera los clústeres se empleó calcular la marca de clase y adicionar la relación entre cantidad de combustible en galones sobre la capacidad instalada, así como también el costo total de traer un galón de diésel a las localidades.

VALORES REALES – MARCAS DE CLASE					
VARIABLES	C1	C2	C3	C4	C5
CAP	56	94	36	26	24
USU	89	186	47	45	32
EFACT	3898	11755	2019	1692	972
CU	1501	1542	1322	1671	1106
FACT	5925945	16692150	2453031	2768553	1382882
SUBS	4867934	14164620	1935868	2356563	1108980
DIE	397	1178	194	198	112
PDIE	6631	6596	6690	6690	10050
PTDIE	520589	1712182	244560	254862	409014
GL/kW	7.09	12.54	5.32	7.64	4.66
CTDIE	7943	8049	7954	7978	13704

Tabla 21: Marcas de clase, de las variables que caracterizan el servicio de prestación de la energía para cada clúster.

A partir de lo anterior, es posible afirmar lo siguiente para cada clúster:

- *Clúster 1 (C1)*: conformado por zonas las cuales su capacidad de generación es media baja, es decir, se encuentra en un rango entre los 20 a los 90 kW. Cuenta con una cantidad de usuarios moderados, el costo unitario es medio, el diésel comprado generalmente opera entre los 300 a los 600 galones. El costo del transporte es medio y cuenta con disponibilidad de combustible considerable.
- *Clúster 2 (C2)*: representa las zonas con mayor capacidad de generación, número de usuarios, energía facturada, y cuentan con bastante disponibilidad de combustible, mayor a 10 GL/kW. Se puede resumir que el C2 es el conglomerado de las zonas grandes.
- *Clúster 3 (C3)*: son localidades pequeñas tanto en capacidad instalada como en número de usuarios, sin embargo, su característica más marcada es que cuentan con valores de disponibilidad de combustible bajos en relación con la capacidad (menor a 6 GL/kW). A pesar de esto, logra generar mucha más energía que otras zonas.
- *Clúster 4 (C4)*: cuenta con zonas que tienen el potencial de generar más energía debido a la disponibilidad de combustible, sin embargo, probablemente no sucede de esa manera debido a que son las localidades que presentan el costo unitario del kWh más alto.
- *Clúster 5 (C5)*: zonas pequeñas en capacidad y número de usuarios, caracterizadas por tener el costo unitario de prestación del servicio más bajo, pero, cuentan a su vez con los precios más altos del galón de diésel, así como el del transporte de este. A tal punto que, el costo de tener un galón de combustible en alguna de esas localidades ronda los 13.704 COP/GL. Presentando un incremento más notorio que en las zonas de los demás clústeres.

## 6.2. CALCULO DE INDICADORES INDIVIDUALES

Se pretende evaluar la viabilidad de implementar gasificación de biomasa en las ZNI que conforman la base de datos mediante el cálculo de indicadores compuestos, sin embargo,

para obtener un indicador compuesto primero es necesario calcular indicadores individuales. Para esto se consideró realizar dos tipos de normalización, la ya mencionada Z-score y la normalización por rango entre 0 y 1. De la cual se decidió emplear la segunda puesto que, se quiere que todos tengan el mismo peso en el indicador compuesto. Los indicadores calculados son los siguientes:

- *nDAU*: es el resultado de normalizar los valores del déficit de atención a usuarios a partir de los valores mínimos y máximos.
- *nDD*: es el resultado de normalizar los valores del déficit de diésel a partir de los valores mínimos y máximos.
- *nDBM*: es el resultado de normalizar los valores del margen de biomasa seca, empleando como máximo el valor de 3, debido a que a pesar de que cambien de manera positiva las condiciones de generación de las zonas, es decir, que empiecen a generar más energía éstas no logran aumentar su generación más de 3 veces.
- *nWBM*: es el resultado de normalizar los valores del margen de biomasa húmeda, empleando como máximo el valor de 3, debido a que a pesar de que cambien de manera positiva las condiciones de generación de las zonas, es decir, que empiecen a generar más energía éstas no logran aumentar su generación más de 3 veces.
- *nTBM*: es el resultado de normalizar los valores de margen total de biomasa, margen que es calculado a partir de la suma de los márgenes de biomasa seca y húmeda.

### 6.3. CALCULO DE INDICADORES COMPUESTOS

La construcción de los indicadores compuestos va a representar la oportunidad de definir en qué zonas resulta viable la implantación del sistema de generación a partir de la gasificación de la biomasa. Para esto es necesario tener en cuenta las tipologías anteriormente establecidas. Por consiguiente, se calcularon una serie de coeficientes para poder ponderar los indicadores individuales y de esta manera encontrar el valor del indicador compuesto que se expresa de la siguiente manera:

$$CI = [\alpha DAU_{norm} + (\beta + \delta) DD_{norm}] * 0.5 + (\gamma BM_{norm}) * 0.5 \quad (6.2)$$

Donde:

$\alpha, \beta, \delta, \gamma$  son coeficientes de ponderación para cada uno de los valores normalizados de los indicadores individuales.

La definición de los coeficientes de ponderación se muestra a continuación:

- *Alpha ( $\alpha$ )*: con este se busca que las zonas que cuentan con más energía facturada sean priorizadas, pues un *DAU* del -10% no representa la misma cantidad de usuarios en una zona que factura más energía que una que factura menos. Por lo cual se le asigna al coeficiente un valor de 0 para la zona que menos energía factura y 1 a la que más lo hace.

- *Betha* ( $\beta$ ): este prioriza aquellas zonas en las cuales el costo total del diésel (precio del galón + transporte por galón) es mayor, asignando 1 a la zona de valor más alto y 0 a la de menor, del parámetro mencionado.
- *Delta* ( $\delta$ ): al igual que con Alpha este pondera con un valor más alto las zonas que cuentan con mayor disponibilidad de combustible para generar (GL/kW). Se asigna 0 y 1 de la misma manera.
- *Gamma* ( $\gamma$ ): como los márgenes de biomasa, no son calculados a partir de las tipologías, ni fueron incluidos en el procedimiento de PCA y clusterización. Entonces se da como un término independiente, entonces 1 es el valor por defecto para cada zona.

Como se observa en la expresión 6.2 se le da igual peso tanto a la suma de los déficits como al margen de biomasa, cada termino se multiplica por 0.5. Lo anterior estipula lo que para el análisis de sensibilidad será el caso base.

El indicador compuesto *nCITB* se calcula a partir del *nDAU*, *nDD* y *nTBM* aplicando la expresión 6.2. El comportamiento del *nCITB* se muestra en la figura 21, en la cual se observa una similitud en los resultados de los indicadores para las zonas que conforman el C1, C3 y C4. Mientras que el C2 y C5 también se comportan de manera similar. Es posible que ese comportamiento se de en base a la cantidad de zonas. Un aspecto importante de la gráfica es que se puede afirmar que los valores altos del indicador empiezan a partir de superar el 0.4. Así mismo, los valores bajos del indicador por lo visto están por debajo del 0.2.

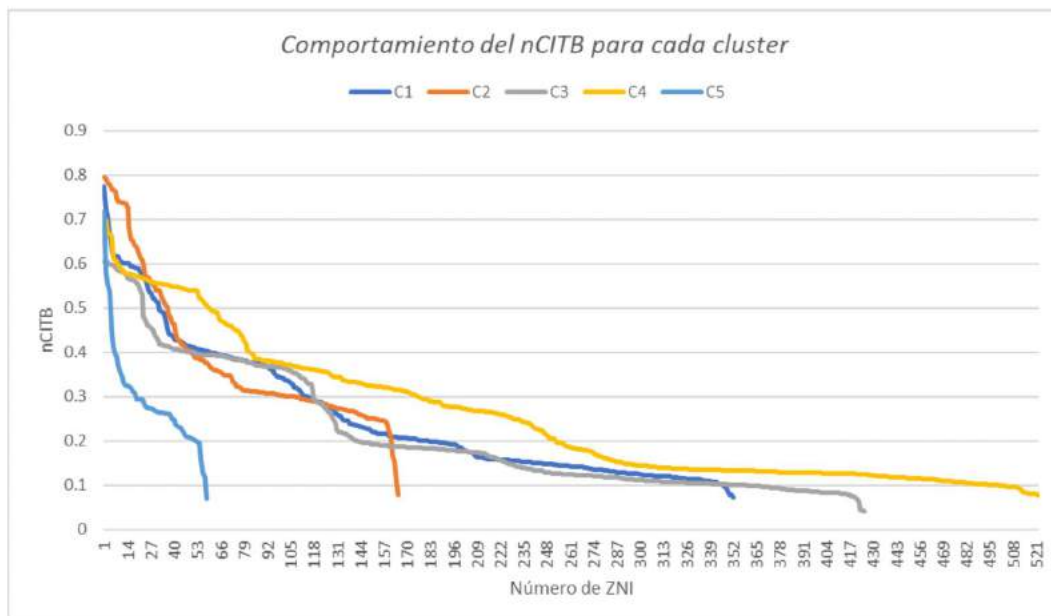


Figura 21: Comportamiento del *nCITB* para cada clúster.

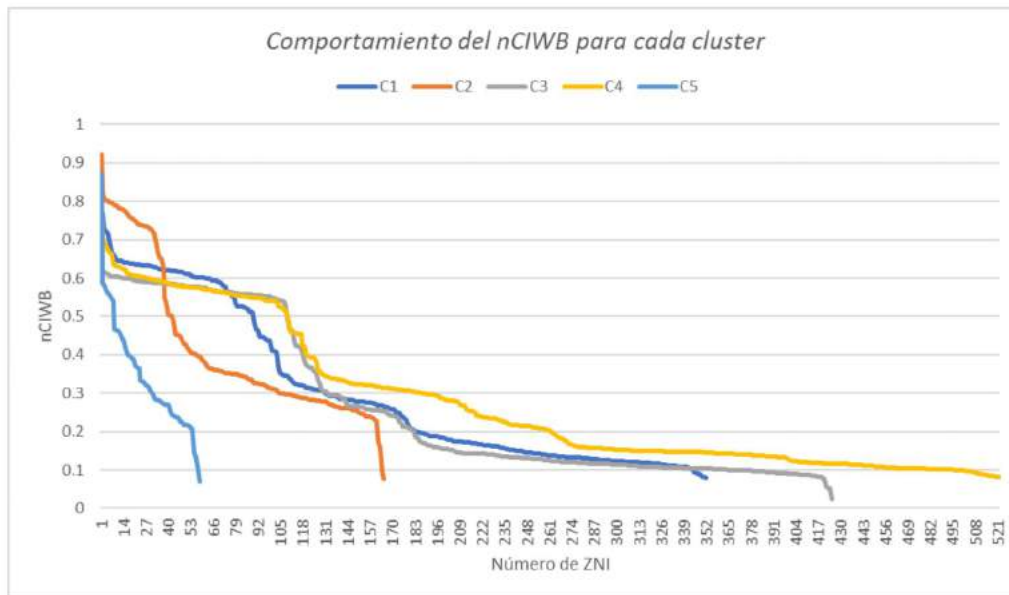


Figura 22: Comportamiento del nCIWB para cada clúster.

El comportamiento del indicador compuesto en base a la biomasa húmeda no representa mayores cambios comparado con el indicador compuesto en base a la biomasa total (Figura 21).

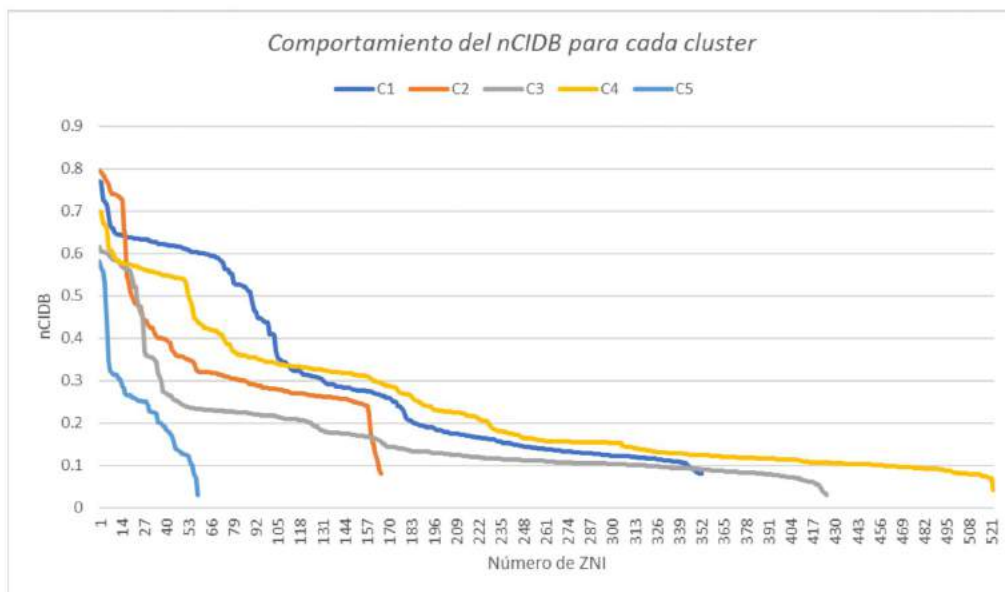


Figura 23: Comportamiento del nCIDB para cada clúster.

En el caso del comportamiento del indicador compuesto en base seca se logra observar, que solo teniendo en cuenta la biomasa del maíz, arroz, palma de aceite y de la caña panelera los valores del indicador compuesto para los clústeres 3 y 4 se corren levemente más a la izquierda, es decir, menos zonas cuentan con un valor más alto. Pasaron de tener aproximadamente 100 zonas por encima de 0.5 en el nCITB a tener menos de 70 en el nCIDB.



#### 6.4. INTERPRETACIÓN DEL INDICADOR COMPUESTO PARA EVALUAR LA FACTIBILIDAD DEL SISTEMA DE GASIFICACIÓN DE BIOMASA

Como ya se mencionó y se observa en las figuras 21, 22 y 23 es posible afirmar que los valores altos de los indicadores compuestos se producen a partir de 0.4 y estos demuestran factibilidad de implementar el sistema de gasificación mientras que los bajos menores a 0.2 corresponden con una NO factibilidad de la implementación. Para comprobar lo dicho anteriormente, se analizó los aportes de cada uno de los términos de la expresión 6.2 para ver cómo están conformados dichos indicadores en los valores altos y bajos de los mismos. Este procedimiento se demuestra con las siguientes figuras:

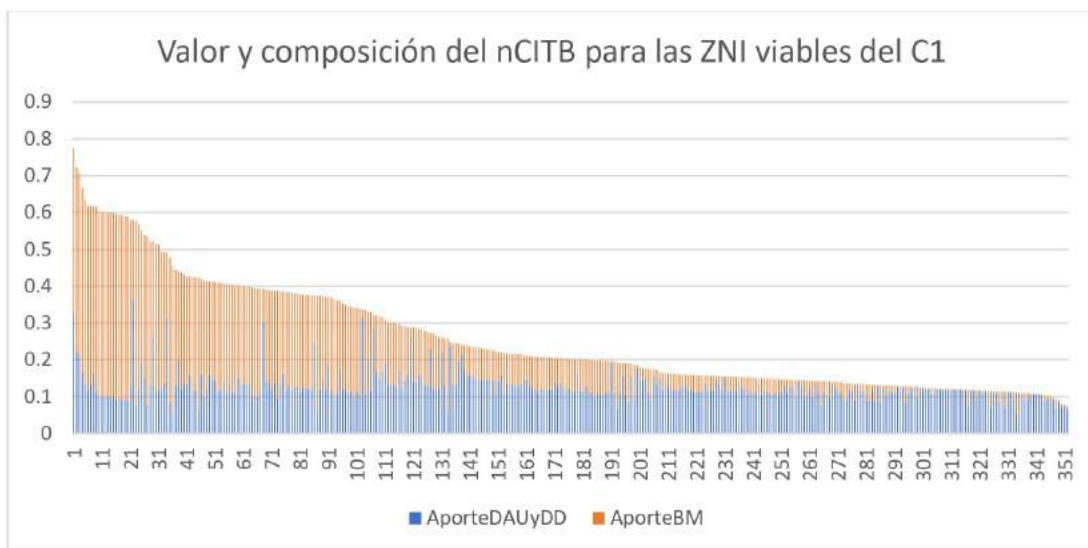


Figura 24: Valor y composición del nCITB para las ZNI del C1.

La figura anterior muestra como está conformada el área bajo la curva del comportamiento del indicador compuesto (figura 21) para la biomasa total de cada zona (nCITB) del clúster 1. Siendo el aporte entregado por la biomasa la parte naranja y el aporte dado por los déficits la zona azul. Para determinar el valor del nCITB a partir del cual la gasificación de biomasa se puede considerar es necesario aclarar que el aporte de biomasa debe ser mayor o igual a **0.08333** debido a que este valor representa el aporte de biomasa en las zonas que pueden generar por lo menos la energía suficiente para suplir su consumo de subsistencia, es decir, por lo menos tener un margen de biomasa total igual a 1. Por consiguiente, el valor del indicador compuesto corresponde con el criterio anterior a partir de 0.2 aproximadamente. Es decir, si el indicador es mayor representa que es factible la generación a partir de gasificación de biomasa en dichas zonas. No obstante, este valor cambia dependiendo del clúster como se ve en las demás figuras.

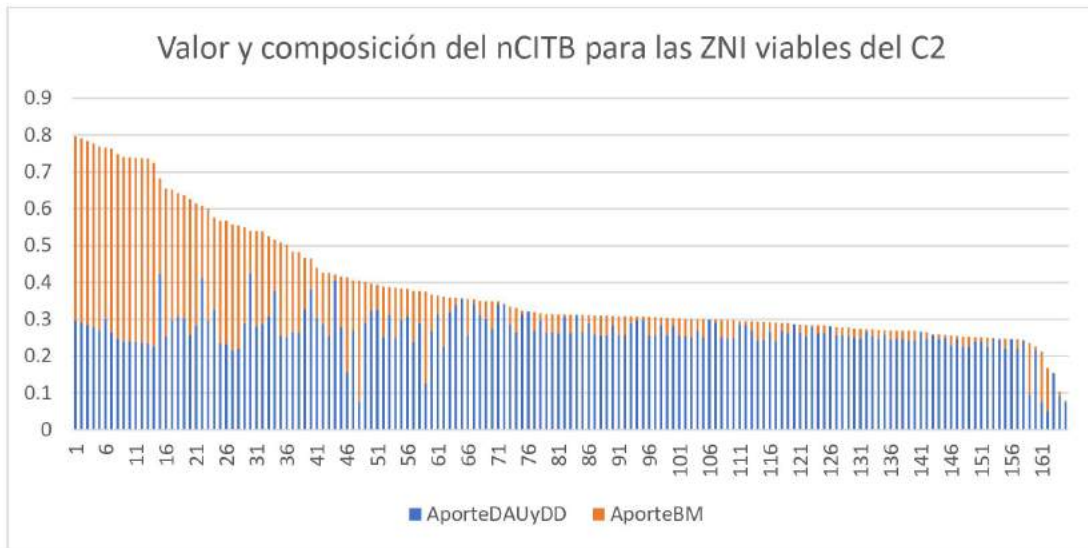


Figura 25: Valor y composición del nCITB para las ZNI del C2.

Para el caso del clúster 2, como se observa el valor del indicador para el cual es factible la gasificación de biomasa se da a partir de aproximadamente 0.35. De ahí para atrás se encuentran zonas que no lograrían suplir completamente su energía de diseño.

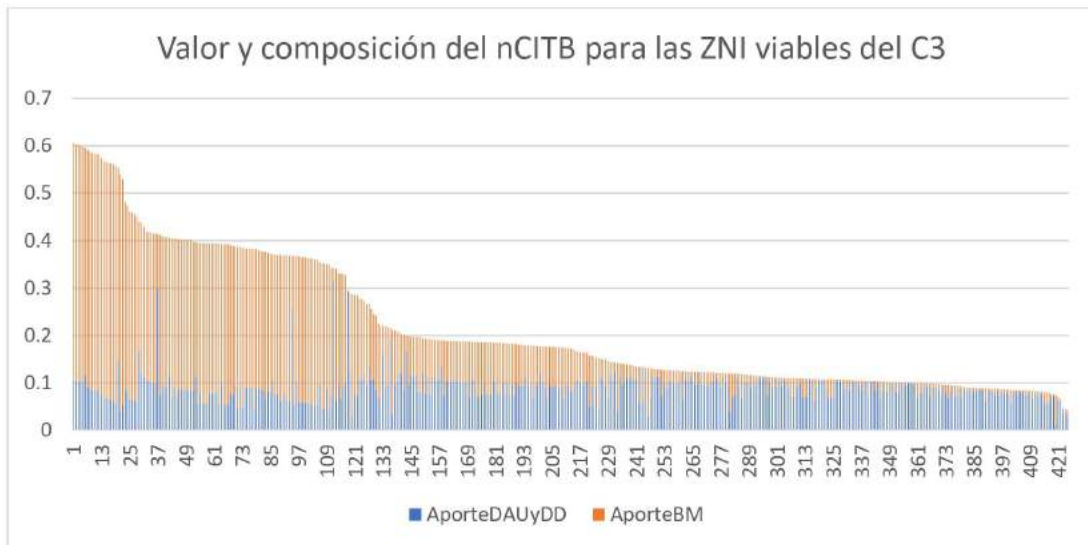


Figura 26: Valor y composición del nCITB para las ZNI del C3.

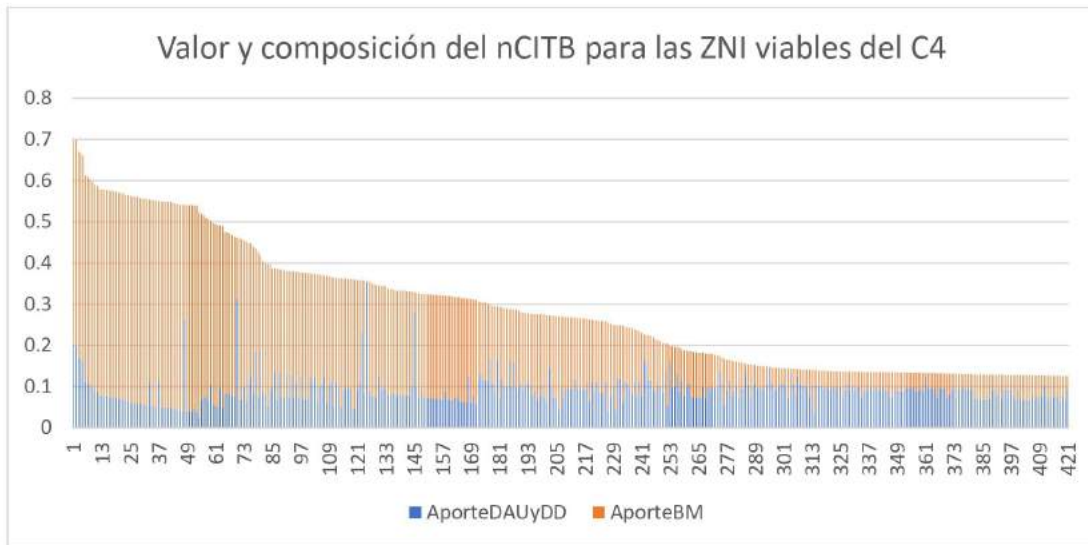


Figura 27: Valor y composición del nCITB para las ZNI del C4.

Los clústeres 3 y 4 representan un comportamiento similar, y es posible definir un valor igual para ambos, valor a partir del cual el indicador compuesto representa un visto positivo para la implementación del sistema de gasificación. Se da a partir de aproximadamente 0.15, la diferencia entre estos clústeres es que las zonas que conforman el C3 y que cuentan con un indicador menor al 0.15, este está compuesto prácticamente por el aporte de los déficits, es decir, estas zonas no cuentan con biomasa. A diferencia del C4 que si tienen un mayor valor en el aporte de biomasa, pero no es suficiente para alcanzar el margen de biomasa total igual a 1.

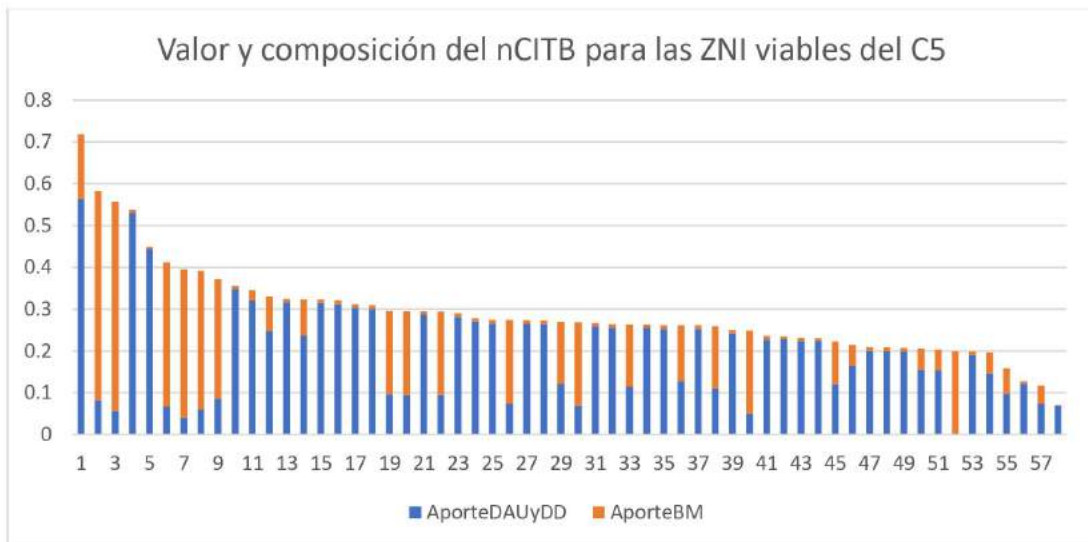


Figura 28: Valor y composición del nCITB para las ZNI del C5.

En el clúster 5, como se observa, no es posible definir un valor del indicador compuesto que represente factibilidad en la implementación de gasificación de biomasa. Principalmente

debido a que los valores de *betha* para este clúster son para la mayoría de zonas muy altos, puesto que, este clúster se caracteriza por el alto costo total del diésel. Sin embargo, al ser pocas zonas es posible analizarlas individualmente, por lo cual se sabe que 21 de ellas tienen factibilidad de implementar el sistema. La distribución del número de zonas en donde el sistema de gasificación es factible se muestra a continuación:

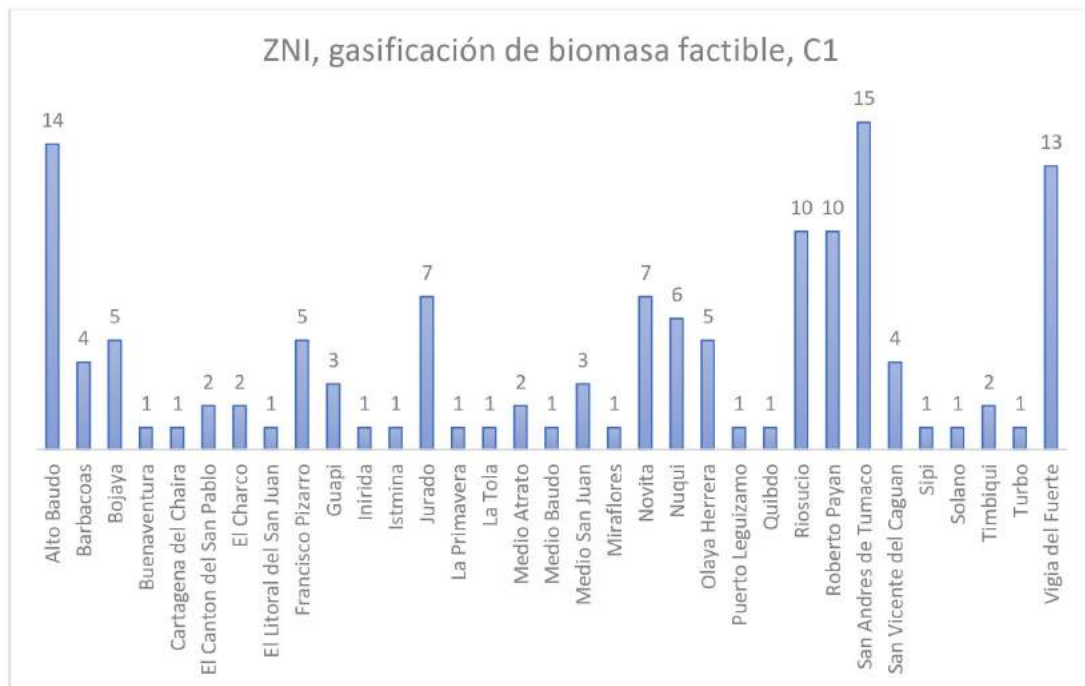


Figura 29: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C1, discriminadas por municipio.

Para el C1 se determinó haciendo una observación más detallada, que el nCITB es más preciso para indicar la factibilidad a partir del 0.25. Las zonas que son iguales o superan dicho valor se muestran en la anterior gráfica discriminadas por el municipio, en total son 133 de las 352 que conforman el C1, es decir, un 37.8% de las zonas tienen factibilidad de implementar el sistema de gasificación para generar energía eléctrica.

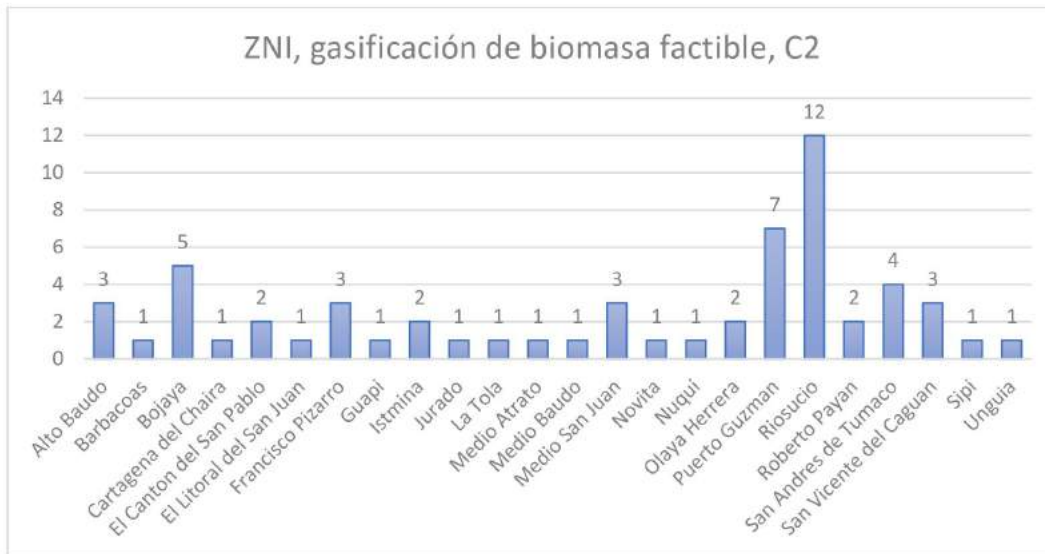


Figura 30: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C2, discriminadas por municipio.

Para el C2 se comprobó lo anteriormente dicho, a partir del 0.35 el nCITB indica la factibilidad de implementar la energización de las zonas en base a gasificación de biomasa. En total, las zonas en donde es factible dicho sistema son 55 de las 165 que conforman el C2, es decir, un 33.33% de las zonas tienen factibilidad de implementar el sistema de gasificación para generar energía eléctrica.

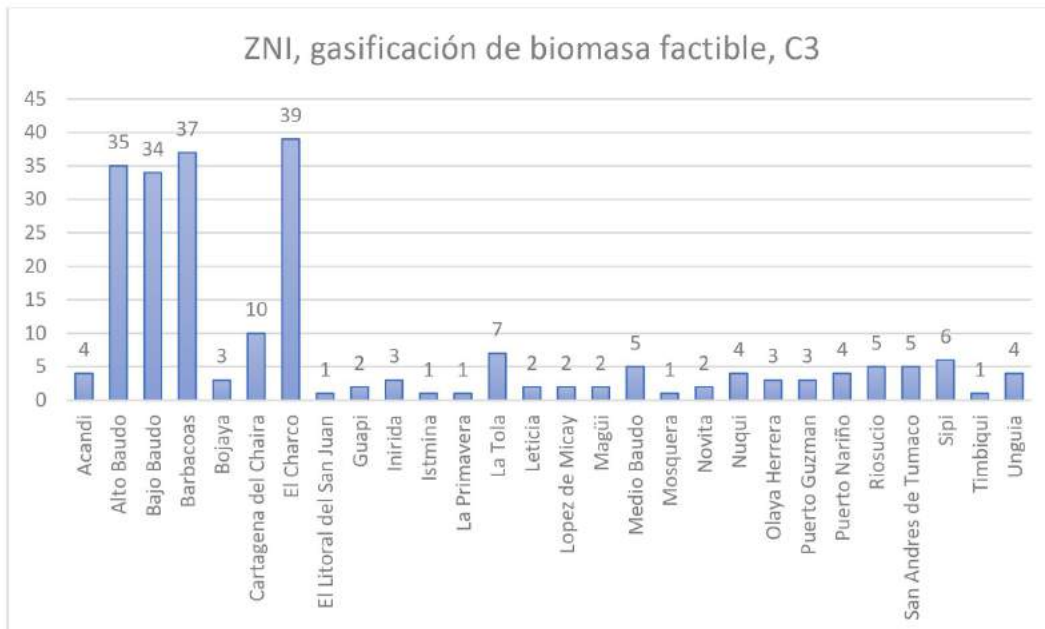


Figura 31: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C3, discriminadas por municipio.

Del clúster 3 se tiene alta concentración de localidades, en los municipios de Alto y Bajo Baudó, Barbaacoas y El Charco. Estos municipios tienen un número de localidades en donde es factible el sistema 145 de las 226. Por otra parte, las zonas en donde es factible son del 53.17% de las 522 que conforman el C3.

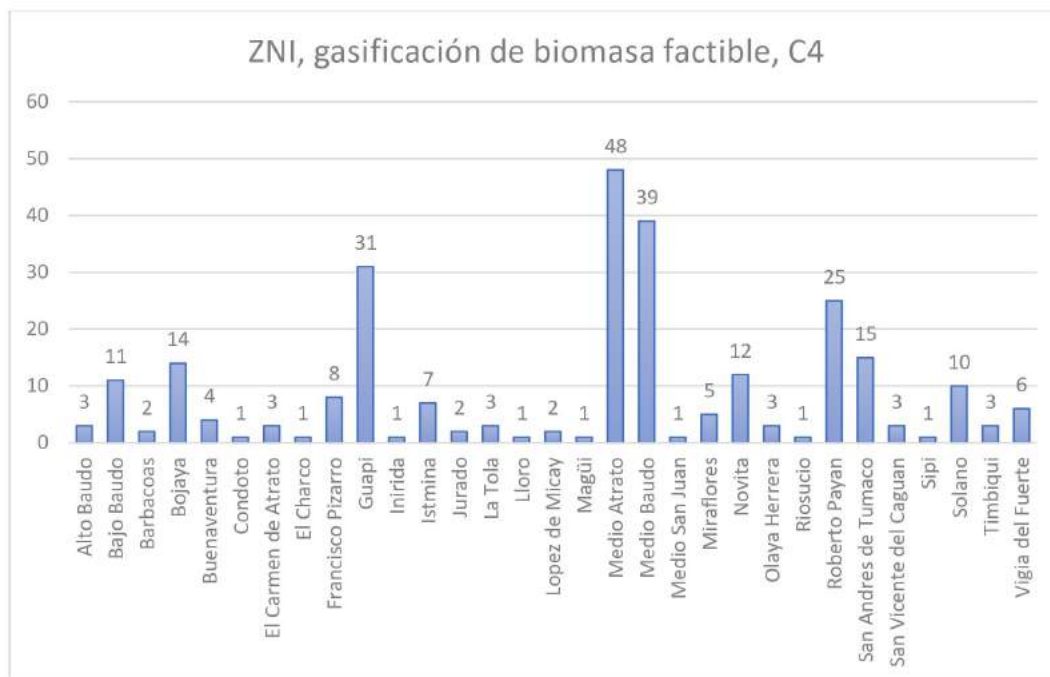


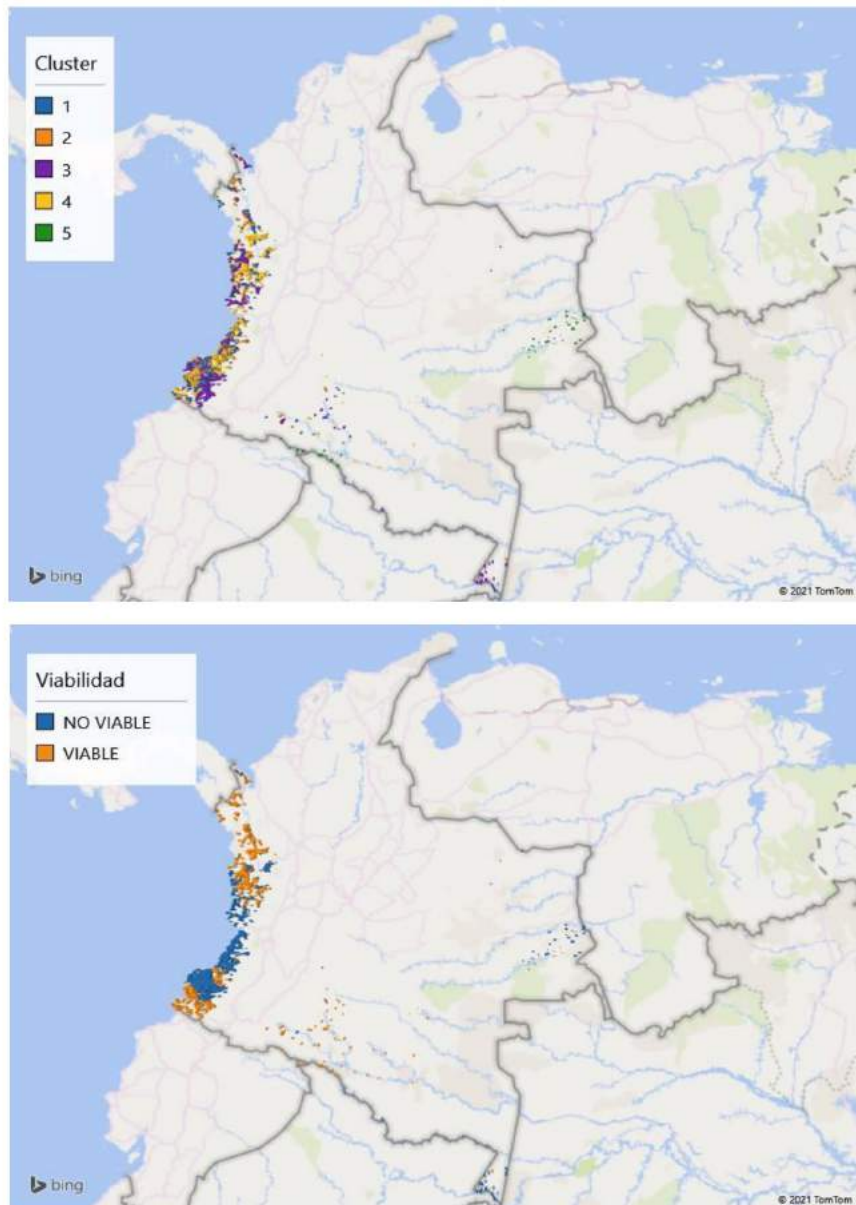
Figura 32: Número de ZNI con posibilidad de generar su energía a partir de la gasificación de biomasa del C4, discriminadas por municipio.

Para el C4 se determinó haciendo una observación más detallada, que el nCITB es más preciso para indicar la factibilidad a partir del 0.18, levemente más alto que el empleado para el C3. En total son 267 de las 522 que conforman el C4, es decir, un 51.1% de las zonas tienen factibilidad de implementar el sistema de gasificación para generar energía eléctrica. Como se muestra en la figura anterior, una gran cantidad de zonas se abarca en los municipios de Guapi, Medio Atrato y Baudó y Roberto Payán. Por otra parte, como ya se mencionó el C5 cuenta con 21 zonas en donde es viable la implementación. En donde, 15 de ellas hacen parte del municipio de Puerto Leguizamo.

De lo anterior, se concluye que a partir del clúster y su tipología correspondiente se definen los valores del indicador de la biomasa total a partir del cual el sistema de gasificación de biomasa es considerado para prestar el servicio de energía eléctrica a las zonas en cuestión. De igual manera, la interpretación de los indicadores que se le da puede tener los siguientes resultados:

- En las zonas en donde el aporte de los dos términos (déficits y margen) sea significativo, entonces en esas localidades será muy pertinente la posibilidad de energizar con base en gasificación de biomasa.

- En donde solo el aporte del margen sea significativo (no hay déficits), esas localidades tienen potencial de usar la gasificación para energizar, pero pueden ser ubicadas en un segundo nivel de pertinencia para su energización con base en gasificación de biomasa porque aún no tienen problemas significativos de suministro.
- En donde el aporte del déficit sea o no significativo y el MB sea muy bajo o nulo. En esas localidades no es posible plantear una solución de energización con base en gasificación de biomasa.



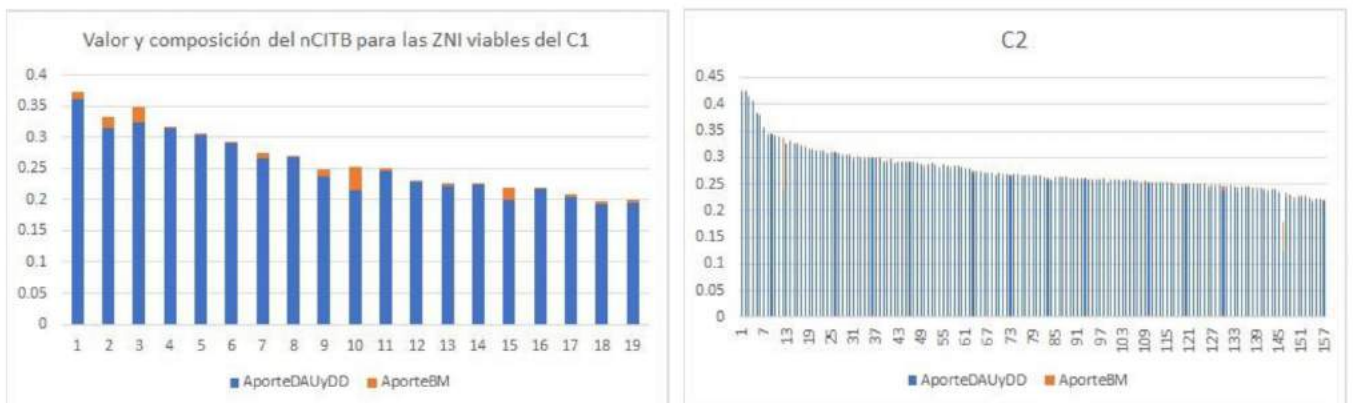
*Figura 33: Visualización de la ubicación geográfica de las ZNI que conforman cada clúster y viabilidad del sistema de gasificación.*

## 6.5. ANALISIS DE SENSIBILIDAD

Los factores de ponderación para el margen de biomasa y para la suma del *DD* y *DAU* se definieron a partir de la lógica del estudio, se busca orientar la investigación a la implementación de gasificación de biomasa como solución energética para dichas zonas. Por consiguiente, se le da mayor peso al margen de biomasa, puesto que, si no se tiene un margen de biomasa considerable da igual que tenga o no la zona *DD* y *DAU* negativo o positivo. Sin embargo, es necesario conocer que tanto afecta la utilización de otros valores para el peso.

Por lo cual se propone establecer como caso base, la configuración anteriormente utilizada, 0.5 tanto para el *nTBM* como para la suma del *DD* y *DAU*. Al cambiar el peso de cada término lo único que se encontró es que los indicadores tienen diferentes aportes, sin embargo, esto no facilita la interpretación del indicador a partir de lo definido en el anterior ítem, puesto que, al cambiar el aporte, es necesario calcular nuevamente, desde que valor del indicador se empieza a considerar la gasificación de biomasa como solución energética para las zonas. Por otra parte, se realizó una variación en el valor de gamma, no se usó el valor por defecto 1 detallado anteriormente, sino se define a partir de que para las localidades más pequeñas es técnicamente más fácil implementar el gasificador, por lo cual para las zonas donde la energía de diseño a la menos 1 es menor (zonas más grandes) se le asigna un 0 y un 1 a las que es mayor (zonas más pequeñas).

De esta manera, se encontró que el indicador compuesto fue altamente sensible, hasta el punto en el que los valores de este fueron dados solo por el aporte del *DD* y el *DAU* en su mayoría para las zonas que conforman los C1, C2 y C5. A diferencia de las que conforman los C3 y C4, en donde si se obtuvo un aporte considerable de biomasa, no obstante, con dicha configuración no es posible determinar un rango de valores en los cuales el indicador compuesto represente de manera precisa la factibilidad de implementar gasificación de biomasa como solución energética. Las figuras que sustentan lo anterior son:





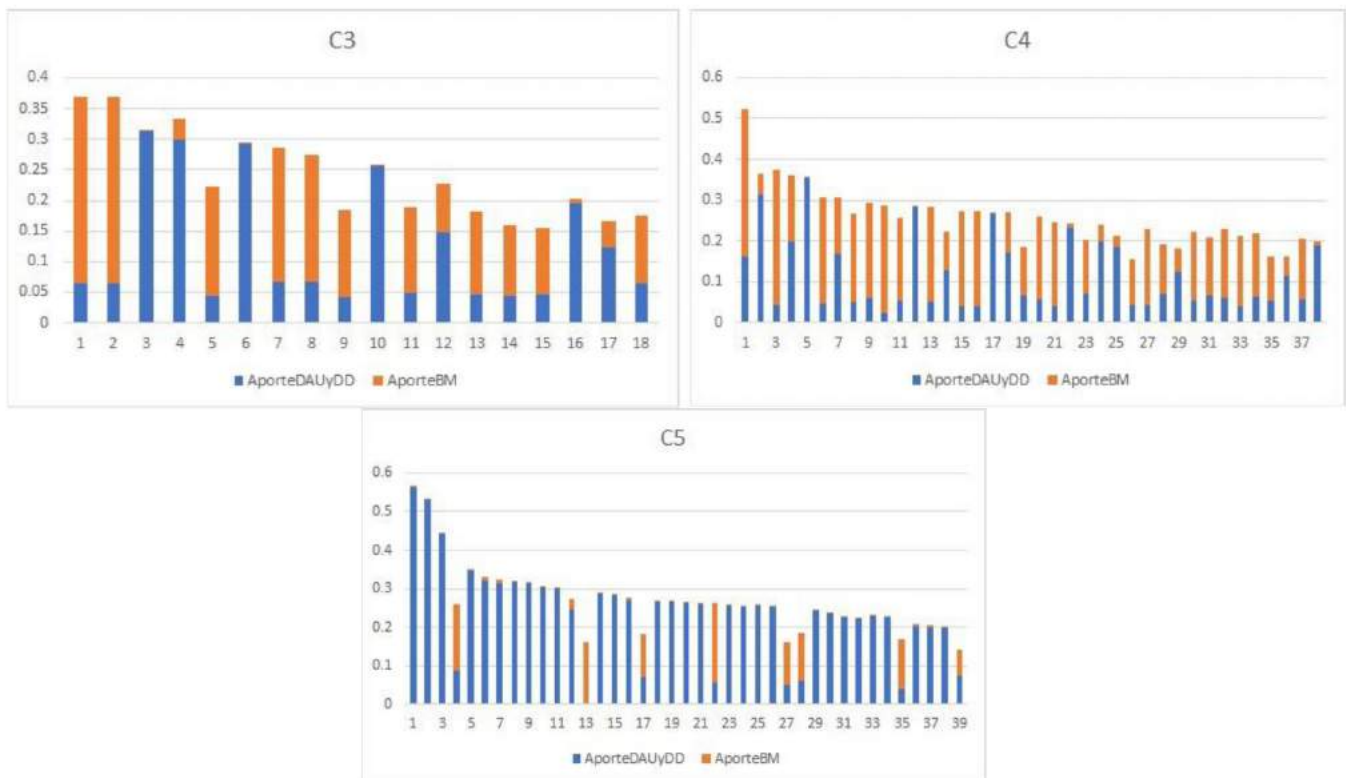


Figura 34: Resultados de los aportes de cada término al indicador global para cada clúster, redefiniendo gamma, coeficiente de ponderación del margen de biomasa.

Con esto se demuestra que la definición de los coeficientes de ponderación representa una de las problemáticas al momento de desarrollar un indicador compuesto, en este caso en específico, haber definido gamma como 1, demostró brindar mejores resultados para el objetivo del proyecto. Pues de esta manera, fue posible a partir de los valores del indicador compuesto y según la tipología a la que pertenece la zona conocer si es factible o si es posible considerar la generación de energía a partir de la gasificación de biomasa.

## 6.6. FACTOR DE CORRECCIÓN DEL INDICADOR COMPUESTO

Como fue posible observar en el ítem 6.4. la interpretación del indicador compuesto está basada en cada uno de los clústeres y para el caso del C5, no es posible determinar un rango de valores en el cual el nCITB represente la factibilidad de implementar el sistema de gasificación como solución energética para las zonas. Con el fin de encontrar un mejor planteamiento del indicador compuesto se realizó el análisis de sensibilidad, sin embargo, los resultados no fueron satisfactorios. No obstante, se analizaron los valores de los aportes tanto de la biomasa como de los déficits y se concluyó que:

Si se realiza un método de ponderación a partir de el producto de los aportes es posible realizar una corrección al indicador compuesto, brindando así que las **zonas que tengan**

**tanto déficit como biomasa alcancen un valor más alto del nCITB.** Esta corrección esta fundamentada en lo siguiente:

- El indicador compuesto obtiene su valor de la suma entre el aporte de biomasa y el de los déficits. El valor total del nCITB para cada zona representa el 100%.
- Se calcula el porcentaje que entregan los aportes al nCITB.
- Se busca priorizar las zonas en donde se tenga tanto biomasa como déficits. Por consiguiente, lo óptimo sería obtener una zona cuyo nCITB esté compuesto por un 50% de cada aporte (50% nTBM y 50% entre nDD y nDAU).
- Al emplear una técnica de agregación mediante el producto de estos porcentajes, es posible asignar un factor de corrección del nCITB, como se ejemplifica a continuación:

Se tiene la localidad de *Limonas* perteneciente al municipio de *Guapi* la cual cuenta con los siguientes indicadores individuales:

Municipio	Localidad	Clúster	nDAU	nDD	nTBM
Guapi	Limonas	1	0.9239406	0.88646685	0.82103068

Tabla 22: Valores de los indicadores individuales para la localidad de *Limonas*, del municipio de *Guapi*.

Por consiguiente, el nCITB da un valor de *0.735064* calculado a partir de la expresión 6.2. teniendo en cuenta los coeficientes de ponderación de la zona. El aporte de los déficits (primer término) es de *0.324549* y el de la biomasa total (segundo término) corresponde a *0.410515*, sumados estos dos, dan como resultado el valor del nCITB ya mencionado. Estos a su vez, representan un *44.15%* y un *55.86%* respectivamente, el producto de estos porcentajes es *0.2465*. Como ya se ha mencionado, se priorizan las zonas con distribución 50/50, este producto es *0.25*, por tanto, el factor de corrección (FC) para esa configuración es de *1*. Mientras que para el caso en cuestión, el FC se calcula a partir de la siguiente expresión:

$$FC = \frac{\text{Producto de los porcentajes de los aportes al nCITB}}{\text{Valor de la configuración deseada (50\% DD y DAU - 50\% BM)}} \quad (6.3)$$

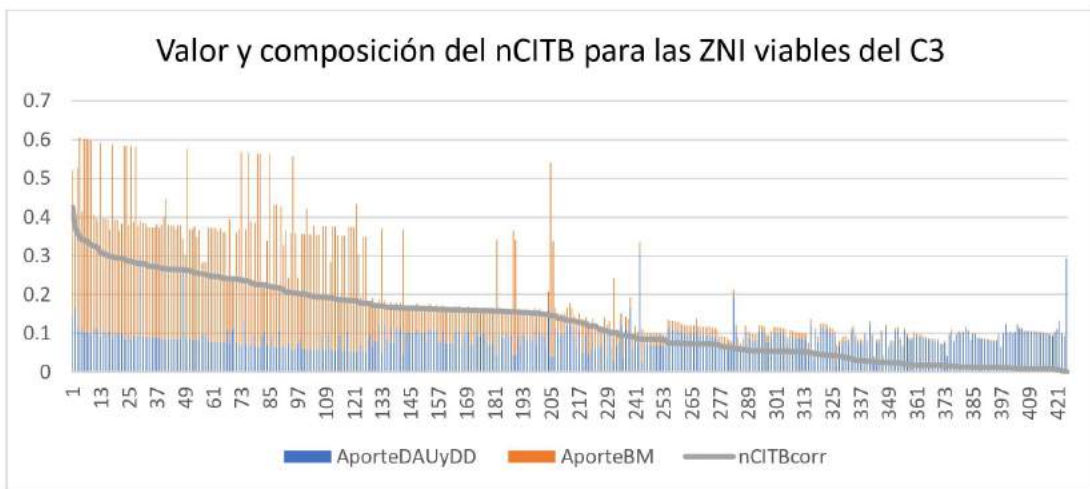
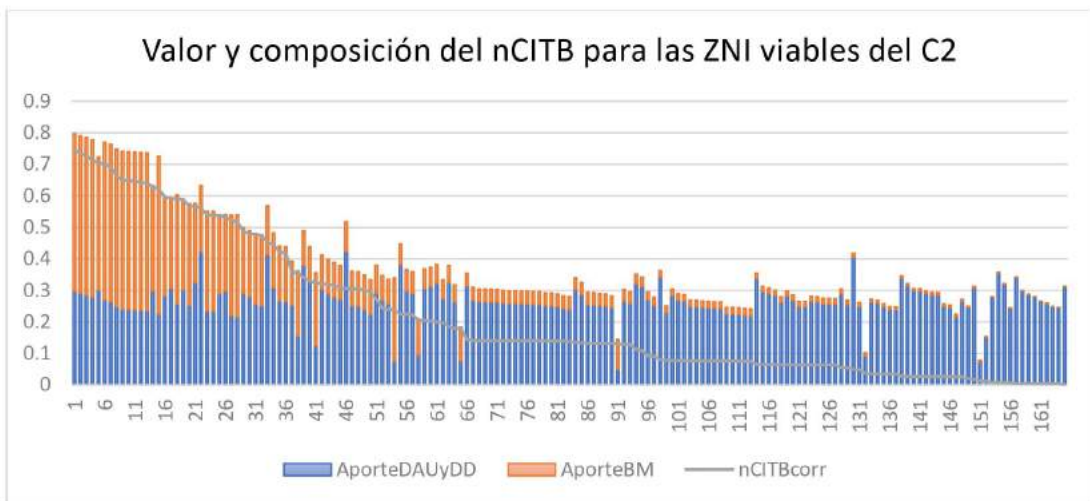
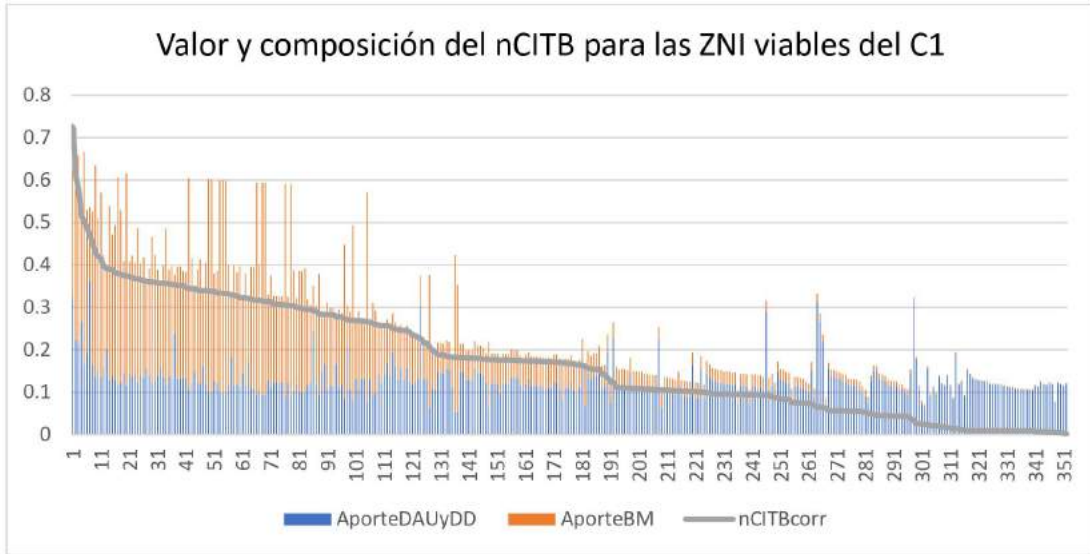
Al emplear la expresión 6.3 en la zona en cuestión se obtiene:

$$FC = \frac{0.2465}{0.25} = 0.9863$$

Por lo cual, el nCITB corregido se expresa:

$$nCITB_{corr} = nCITB * FC \quad (6.4)$$

Entonces, para el caso en cuestión el nCITB corregido es de *0.725*. Los resultados obtenidos en esta configuración son los siguientes para cada uno de los clústeres:



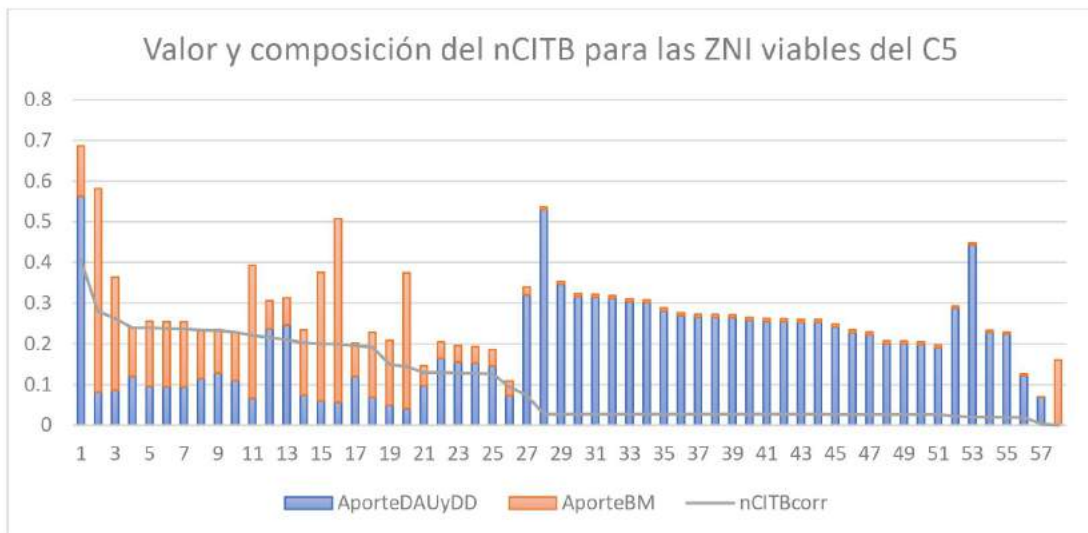
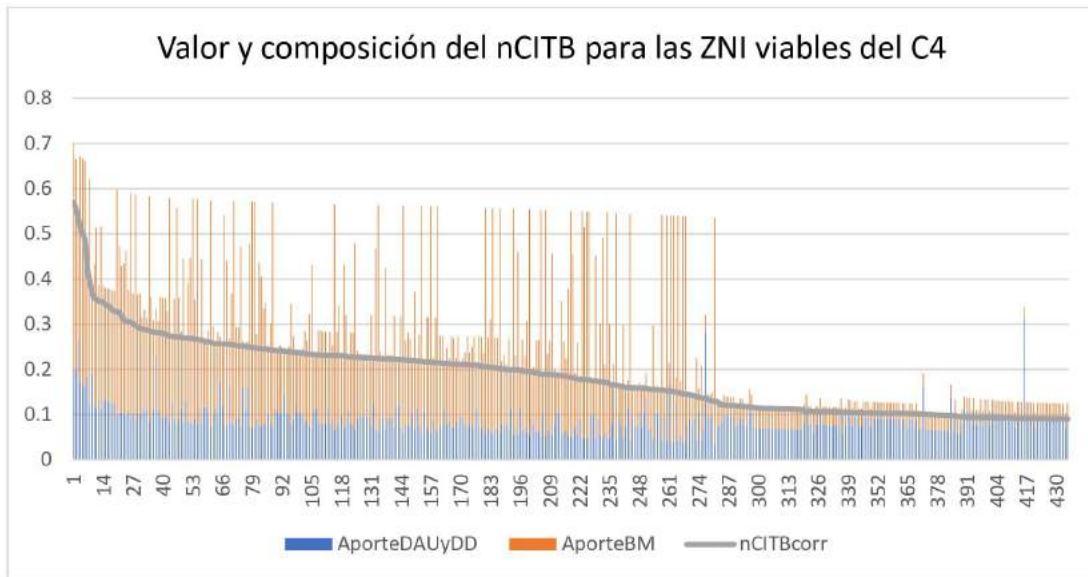


Figura 35: Resultados del indicador compuesto corregido ( $nCITB_{corr}$ ) comparado con los aportes de los déficits y biomasa del indicador base global ( $nCITB$ ) para cada cluster.

Con las figuras anteriores es posible observar el efecto que tiene el factor de corrección, generando que los valores más altos del indicador compuesto pertenezcan a zonas que tienen contribución de ambos aportes e influyendo en zonas como las del C5 que cuentan con mayor aporte de los déficits, para estas, el indicador  $nCITB$  cae drásticamente, representando la NO factibilidad de implementar el sistema de gasificación de biomasa. Con esta configuración, es posible también definir un único valor con el cual es posible evaluar la factibilidad independientemente del clúster al que pertenezca la zona. Esto es posible a partir de observar los valores mínimos del indicador corregido para las zonas donde es factible, mencionadas en el ítem 6.4.

- *CLUSTER 1, mínimo del nCITBcorr: 0.1617*
- *CLUSTER 2, mínimo del nCITBcorr: 0.13*
- *CLUSTER 3, mínimo del nCITBcorr: 0.1018*
- *CLUSTER 4, mínimo del nCITBcorr: 0.089*
- *CLUSTER 5, mínimo del nCITBcorr: 0.1439*

De estos valores se conoce que el más grande comparados entre sí es de 0.1617, por lo cual si se emplea este valor como punto de partida para evaluar la factibilidad, este es posible aplicarse para cualquiera de los clústeres. Puesto que, independientemente de la zona, mientras que el nCITB corregido supere dicho valor, indicará que es factible implementar el sistema de gasificación de biomasa. También se estudió que al emplear este no se descartara una cantidad de zonas considerables, mostrando que en el C2 solo una de ellas está por debajo, en el C3 20 zonas, en el C4 21 zonas y por último en el C5 se descartarían 3 zonas. De esta manera, se encuentra que 622 zonas de las 1522 son zonas en donde el sistema de gasificación es factible, a diferencia de las 706 zonas identificadas en la configuración anterior. Aun así, se considera que esta configuración representa de mejor manera lo que se quiere, evaluar la factibilidad del sistema teniendo en cuenta los aportes de los déficits y la disponibilidad de biomasa, siendo evaluado de la misma manera para todas las zonas.

## CONCLUSIONES

Se logró obtener tipologías que caracterizan y diferencian el estado de prestación de servicio para las zonas no interconectadas, gracias a la implementación del algoritmo de clusterización. La importancia de las tipologías radica en que permite evaluar la factibilidad de implementar el sistema de gasificación de biomasa conociendo los valores de las variables reportadas al CNM y el margen de biomasa de una ZNI cualquiera, pues al conocer estos valores se asigna dicha zona a una de las tipologías, se calculan los indicadores individuales y posteriormente el indicador compuesto, comparando si supera o no el valor del indicador compuesto necesario para determinar la viabilidad según la tipología.

Con la evaluación del indicador compuesto, fue posible determinar que el 46.39% de las zonas que cuentan con datos de los cultivos (1522) y que fueron evaluadas según sus aportes tanto del *DD*, *DAU* y *TBM* son factibles para prestar su servicio de generación eléctrica a partir de la gasificación de biomasa, puesto que, cuentan con biomasa suficiente para suplir por lo menos la energía de diseño. Sin tener en cuenta el cluster al que pertenecen, los 5 municipios que más zonas con factibilidad son Alto Baudo, Medio Atrato, Bajo y Medio Baudo, y Barbacoas, con 55, 51, 46, 46 y 44 zonas correspondientemente. Sin embargo, estos resultados se obtienen evaluando la factibilidad en un rango de valores del nCITB que depende de cada clúster.

La implementación de un factor de corrección para el nCITB permite evaluar la factibilidad del sistema de gasificación de biomasa en las zonas sin que sea dependiente del clúster al que pertenezca, se encuentra factible a partir de 0.1617 en el valor del nCITB corregido. Con esta configuración se encuentra que 622 de las zonas son factibles. Esto representa un 37.78%, 32.73%, 32.71%, 45.21% y 25,86% de las zonas correspondientes a los C1, C2, C3, C4 y C5 respectivamente.

Durante el proceso de análisis de componentes principales (PCA) se pudo reunir más del 90% de la varianza de todos los datos en solo 3 componentes. Un valor elevado que se asocia a la precisión con la que fueron presentados las ZNI como un punto en específico en los diagramas de dispersión, a su vez esto brinda mayor acierto para el proceso de clusterización y si se quiere, es posible obtener los valores reales a partir de las PC's y corroborar el error asociado a ese 9% de varianza perdida.

Se desarrollo una metodología de pronóstico para el potencial energético de la biomasa para el periodo en cuestión, aplicable para cualquier otro periodo superior a 2018, tenido en cuenta, los rendimientos promedios y máximos que han tenido los cultivos en Colombia desde 2007. A su vez, los potenciales técnicos de biomasa se calcularon teniendo en cuenta una eficiencia de conversión eléctrica del 13% valor típico de un gasificador – motor que opera a partir de biomasa secas y un factor de recolección de residuos de 0.05 para la caña panelera y de 0.8 para el resto de los cultivos en cuestión.

Se pudo comprobar que el comportamiento de la prestación del servicio de energía eléctrica no sufre cambios drásticos en cuestión del déficit de atención a usuarios y de diésel. Puesto que, las desviaciones estándares no superan más de un 30% porcentual para la mayoría de las zonas. Por lo cual, se puede afirmar que las zonas que han presentado diésel durante un periodo de tiempo, está situación no cambia de manera drástica, es decir, sigue conservando el déficit solo con pequeñas variaciones en su magnitud.

La implementación del algoritmo de clusterización con ( $k=5$ ) fue posible gracias a la remoción de outliers, puesto que, de lo contrario los clústeres se distribuían de manera muy ineficiente, a tal punto que uno de ellos era conformado por solo 7 localidades como se observa en la figura 14.

## RECOMENDACIONES

Se aconseja emplear otros parámetros de evaluación de evaluación del número de clústeres óptimo, en donde, sea más clara la determinación. En este caso, fue necesario definirlo a partir de método gráfico. Lo más recomendable sería definir  $k$  a partir de criterios netamente teóricos.

Para las zonas que conforman el clúster 1 y 2, que por tipología son más grandes en términos de número de usuarios y energía facturada, se recomienda que se evalúen los indicadores compuestos de manera menos restringida, que se considere factible la generación con márgenes menores a la unidad. Debido a que, posiblemente se habla de una cantidad de energía mucho mayor que es posible generar a partir de la gasificación de biomasa que la generada en una zona que pertenezca a los otros clústeres.

Vista la cantidad de zonas por municipio, se recomienda para trabajos futuros, tomarlas como referencia para diseñar un sistema de gasificación total que logre generar la energía de las zonas de esos municipios. De esta manera, es probable que se pueda suministrar el servicio a todas las zonas, incluso las que no tengan la biomasa suficiente. Debido a que algunas zonas tienen márgenes de biomasa mucho mayores a la unidad, lo que significa que pueden llegar a generar más energía de la de diseño, creando la oportunidad prestarle el servicio a zonas que no cuentan con la biomasa.



## BIBLIOGRAFIA

- Bello Vasquez, D., & Caicedo Medina, L. E. (1996). *Manual para cálculo y diseño de un trapiche panelero*. Tesis de pregrado, Corporación Universitaria Autónoma de Occidente, División de Ingenierías, Santiago de Cali. Obtenido de <https://red.uao.edu.co/bitstream/handle/10614/3473/T0001363.pdf;jsessionid=697C4225942C72F4F6932BB5B45B195A?sequence=1>
- Bocci, E., Moneti, M., Vecchione, L., Di Carlo, A., Villarini, M., & Sisinni, M. (2014). State of art of small scale biomass gasification power systems: a review of the different typologies. (Elsevier, Ed.) *Energy Procedia*, 45, 247-256.
- Centro Internacional de Agricultura Tropical (CIAT), Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT). (2019). *Maíz para Colombia visión 2030*. Obtenido de <https://www.fenalce.org/archivos/maiz2030.pdf>
- Cicea, C., Marinescu, C., & Pintilie, N. (2020). New Methodological Approach for Performance Assessment in the Bioenergy Field. (S. I. Mussatto, Ed.) *Energies*. doi:<https://doi.org/10.3390/en14040901>
- Endel, F., & Piringer, H. (2015). Data Wrangling: Making data useful again. (Elsevier, Ed.) *IFAC-PapersOnLine*, 48, 111-112. doi:<https://doi.org/10.1016/j.ifacol.2015.05.197>
- FINAGRO. (2021). *Sección de Noticias de FINAGRO*. Obtenido de Los secretos ocultos de la biomasa, gran potencial que posee Colombia: <https://www.finagro.com.co/noticias/los-secretos-ocultos-de-la-biomasa-gran-potencial-que-posee-colombia>
- Flórez Acosta, J. H., Tobón Orozco, D., & Castillo Quintero, G. A. (2009). ¿Ha sido efectiva la promoción de soluciones energéticas en las ZNI en Colombia?: Un análisis de la estructura institucional. *Cuaderno de Administración*, 219-245.
- Gorgas, J., & Cardiel, N. (2011). *Análisis de datos y Estadística Avanzada, PCA*. Universidad Complutense de Madrid, Facultad de Ciencias Físicas, Madrid. Obtenido de [http://webs.ucm.es/info/Astrof/POPIA/asignaturas/ana\\_dat\\_est/tema09\\_x2.pdf](http://webs.ucm.es/info/Astrof/POPIA/asignaturas/ana_dat_est/tema09_x2.pdf)
- Grupo Bancolombia. (6 de Septiembre de 2018). *El ABC de la producción de arroz en Colombia*. Obtenido de <https://www.grupobancolombia.com/wps/portal/negocios/actualizate/sostenibilidad/cultivo-arroz-en-colombia>
- Jenkins, S. (2020). *2019 Chemical Engineering Plant Cost Index Annual Average*. Obtenido de <https://www.chemengonline.com/2019-chemical-engineering-plant-cost-index-annual-average/>
- Jolliffe, I. (2002). *Principal Component Analysis* (Segunda ed.). New York: Springer.

- Karaj, S., Rehl, T., Leis, H., & Muller, J. (Enero de 2010). Analysis of biomass residues potential for electrical energy generation in Albania. *Renewable and Sustainable Energy Reviews*, 14, 493-499. doi:<https://doi.org/10.1016/j.rser.2009.07.026>
- Kaufman, L., & Rousseeuw, P. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Bruselas, Belgica: Wiley Interscience. Obtenido de [http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/503\\_09\\_Kaufman\\_Finding-Groups-in-Data-An-Introduction-to-Cluster-Analysis.pdf](http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/503_09_Kaufman_Finding-Groups-in-Data-An-Introduction-to-Cluster-Analysis.pdf)
- Lu, Y., Guo, L., Ji, C., Zhang, X., Hao, X., & Yan, Q. (Junio de 2006). Hydrogen production by biomass gasification in supercritical water: A parametric study. (Elsevier, Ed.) *International Journal of Hydrogen Energy*, 31, 822-831. doi:<https://doi.org/10.1016/j.ijhydene.2005.08.011>
- Microsoft. (2021). *Soporte de Office*. Obtenido de PRONOSTICO.ETS (función PRONOSTICO.ETS): <https://support.microsoft.com/es-es/office/pronostico-ets-funci%C3%B3n-pronostico-ets-15389b8b-677e-4fbd-bd95-21d464333f41>
- Ministerio de Agricultura. (2014). *Banano, Evaluaciones Agropecuarias Municipales*. Bogotá. Obtenido de <https://www.agronet.gov.co/documents/banano.pdf>
- Ministerio de Agricultura. (2014). *Caña panelera, Evaluaciones Agropecuarias Municipales*. Obtenido de <https://www.agronet.gov.co/Documents/Ca%C3%B1a%20Panelera.pdf>
- Mirkin, B. (2005). *Clustering for Data Mining: A data recovery approach* (Primera ed.). (C. a. Hall/CRC, Ed.) New York: Taylor and Francis Group. doi:<https://doi.org/10.1201/9781420034912>
- Munder, S., Karaj, S., Gummert, M., Haefele, S., & Muller, J. (2012). Improving thermal conversion properties of rice straw by briquetting. *International Workshop on the innovative uses of rice straw and rice husk*.
- Muñoz, A. J. (2003). *Sistema de telemetría y control del servicio de energía eléctrica*. Universidad del Cauca, Popayan. Obtenido de <http://repositorio.unicauca.edu.co:8080/bitstream/handle/123456789/2112/SISTEMA%20DE%20TELEMETRIA%20Y%20TELECONTROL%20DEL%20SERVICIO%20DE%20ENERGIA%20ELECTRICA.pdf?sequence=1&isAllowed=y>
- Murillo, T. (1999). Alternativas de uso para la gallinaza. *XI Congreso Nacional Agronomo Conferencia 94* (págs. 427-436). San José: Pecuarios. Obtenido de [http://www.mag.go.cr/congreso\\_agronomico\\_xi/a50-6907-III\\_427.pdf](http://www.mag.go.cr/congreso_agronomico_xi/a50-6907-III_427.pdf)
- Open Knowledge Foundation. (2013). *Data Wrangling Handbook* (Primera ed.). School Data.
- Organization for Economic Co-Operation and Development (OECD). (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*.

- Peña, D. (2002). *Análisis de datos multivariantes*. Madrid, España. Obtenido de [https://www.researchgate.net/publication/40944325\\_Analisis\\_de\\_Datos\\_Multivariantes](https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes)
- Pérez Saldarriaga, L., & Salazar Briceño, S. (2015). *Elaboración de una guía ambiental para el manejo limpio en el aprovechamiento energético de la biomasa residual*. Universidad de la Salle, Ingeniería Ambiental y Sanitaria, Bogotá. Obtenido de [https://ciencia.lasalle.edu.co/ing\\_ambiental\\_sanitaria/546/?utm\\_source=ciencia.lasalle.edu.co%2Fing\\_ambiental\\_sanitaria%2F546&utm\\_medium=PDF&utm\\_campaign=PDFCoverPages](https://ciencia.lasalle.edu.co/ing_ambiental_sanitaria/546/?utm_source=ciencia.lasalle.edu.co%2Fing_ambiental_sanitaria%2F546&utm_medium=PDF&utm_campaign=PDFCoverPages)
- Pollesch, N. L. (2016). Mathematical Approaches to Sustainability Assessment and Protocol Development for the Bioenergy Sustainability Target Assessment Resource (BioSTAR). *Bioenergy*. Obtenido de [https://trace.tennessee.edu/utk\\_graddiss/3872/](https://trace.tennessee.edu/utk_graddiss/3872/)
- Rodrigo, J. A. (Junio de 2017). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. (A. 4. 4.0), Ed.) Obtenido de Ciencia de datos: [https://www.cienciadedatos.net/documentos/35\\_principal\\_component\\_analysis](https://www.cienciadedatos.net/documentos/35_principal_component_analysis)
- Romero, D. A. (2019). *Determinación del potencial energético de la biomasa residual de cultivos de banano en el Cantón Machala, El Oro, Ecuador*. Tesis de pregrado, Cuenca. Obtenido de <https://dspace.ups.edu.ec/bitstream/123456789/18090/1/UPS-CT008600.pdf>
- Sistema Único de Información Normativa. (18 de Diciembre de 2003). *Artículo 1 de la Ley 855*. Obtenido de <http://www.suin-juriscol.gov.co/viewDocument.asp?id=1669722>
- Software para encuestas, QuestionPro. (2020). *Tamaño de muestra*. (A. Aguilar, Editor) Recuperado el 15 de Mayo de 2021, de <https://www.questionpro.com/es/tama%C3%B1o-de-la-muestra.html>
- Superintendencia de Servicios Públicos Domiciliarios. (Septiembre de 2020). *Informe sectorial de la prestación del servicio de energía eléctrica en las ZNI de Colombia*. Bogotá.
- Talukder, B., Hipel, K., & VanLoon, G. W. (2017). Developing Composite Indicators for Agricultural Sustainability Assessment: Effect of Normalization and Aggregation Techniques. *Resources*. doi:<https://doi.org/10.3390/resources6040066>
- Turton, R., Shaeiwitz, J., Bhattacharyya, D., & Whiting, W. (2018). *Analysis, Synthesis, and Design of Chemical Processes* (Quinta ed.). (Kindle, Ed.) International Series in the Physical and Chemical Engineering Sciences.
- Universidad Distrital Francisco José de Caldas. (2020). *Estado de la cobertura eléctrica y las zonas no interconectadas en la región central*.

University of Minnesota, Department of Computer Science and Engineering. (2014). *Data clustering: Algorithms and applications*. (C. Charu, & K. Chandan, Edits.) Florida, United States: CRC Press.

UPME, IDEAM, COLCIENCIAS, UIS. (s.f.). *Atlas del Potencial Energético de la Biomasa Residual en Colombia*. Obtenido de <https://www1.upme.gov.co/siame/Paginas/atlas-del-potencial-energetico-de-la-biomasa.aspx>

# **ANEXOS**

## CODIGO PARA EL PCA y LA CLUSTERIZACIÓN

### PROCEDIMIENTO GENERAL

```
clc %Borrar el command window
clear all %Borrar todos los datos y variables

%CARGAR LOS DATOS
Sep2020=importfile("G:\Mi
unidad\UNIVERSIDAD\TESIS\PCA\Ejercicios\DatosZNI_Complete.xlsx",
"Sep2020",[2,1590])
%Sep2019=importfile("G:\Mi
unidad\UNIVERSIDAD\TESIS\PCA\Ejercicios\DatosZNI_Complete.xlsx",
"Sep2019",[2,1607])
%Mar2020=importfile("G:\Mi
unidad\UNIVERSIDAD\TESIS\PCA\Ejercicios\DatosZNI_Complete.xlsx",
"Mar2020",[2,1608])
%Mar2019=importfile("G:\Mi
unidad\UNIVERSIDAD\TESIS\PCA\Ejercicios\DatosZNI_Complete.xlsx",
"Mar2019",[2,1553])

%SELECCIONAR EL PERIODO DE TIEMPO
Matriz=Sep2020;
%Construir tabla
%Asignar variables categoricas
CatVar=Matriz(:,1:8);
TYPEZNI=CatVar(:,6);
TYPEZNI=table2array(TYPEZNI);
%Asignar valores de las variables que serán usadas en el PCA
Data=[Matriz(:,9:10),Matriz(:,12:13),Matriz(:,15:18),Matriz(:,20:21)]
Data(:,2)=[];
Data=table2array(Data);
```

### ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

```
%PCA
%Centralizar las variables
meanofdata=mean(Data);
desv=std(Data);
[z,v]=size(Data);
datacentralized=zeros(z,v); %Centralized and standarized

for i=1:v
    for j=1:z
        datacentralized(j,i)=(Data(j,i)-meanofdata(i))/desv(i);
        % datacentralized(j,i)=(Data(j,i)-meanofdata(i));
    end
end
```

```

end

%Comprobar centralización  $\mu=0$  ,  $\sigma=1$ 
meanprueba=mean(datacentralized)
desvprueba=std(datacentralized)

%Uso de matriz de covarianza o matriz de correlación
cov_matriz=cov(datacentralized)
corr_matriz=corrcoef(datacentralized)
a=cov_matriz;
b=corr_matriz;

%Extracción de eigenvectors y eigenvalues
[evect,eval]=eig(a)

%Calculo de los scores (valores de cada PC para las ZNI)
scores=evect'*datacentralized'

%Diagrama de pareto para ver número óptimo de PCs
pareto(diag(eval))

%Probar recuperar datos, procedimiento para verificar la ejecución del PCA
datastorecent=eval*scores;
datastore=datastorecent'+meanofdata(1)
%Si al realizarse, se obtienen los datos originales quedó bien ejecutado

%Calculo de la varianza explicada
scores=scores';
varianza=zeros(v);
for i=1:v
    varianza(i)=(std(scores(:,i)))^2;
end
totvar=sum(varianza);
for i=1:v
    explainedcod(i)=varianza(i)/totvar(1);
end
[explainedcod, idx]=sort(explainedcod,"descend");
pareto(explainedcod)
%varianza explicada equivale a los eval, comprobación

%Ordenar los PCs desde el más influyente al menos
idx=v;
scorescod=scores;

```

```

coeffcod=evec;
for i=1:v
    scorescod(:,i)=scores(:,idx);
    coeffcod(:,i)=evec(:,idx);
    idx=idx-1;
end

%¿Qué resultados se obtienen con la función de matlab?

%con data centralizada
[coeffcent,scorecent,~,~,explainedcent,~] = pca(datacentralized);
scorescod=scorecent;

%Vista tridimensional y vistas 2D
scatter3(scorecent(:,1),scorecent(:,2),scorecent(:,3),10)
scatter(scorecent(:,1),scorecent(:,2),10)
scatter(scorecent(:,1),scorecent(:,3),10)
scatter(scorecent(:,2),scorecent(:,3),10)

%BIPLOT
TagsPCA=["Cap" , "UsuaE1" , "EfacE1" , "CU" , "FacE1" , "SubsE1" , "Diecomp" ,
"PrecGL" , "CostTrans"]
biplot(coeffcod(:,1:3), "Scores", scorescod(:,1:3), "VarLabels", TagsPCA)

%Histograma para el PC1
subplot(2,2,1);
histogram(scorescod(:,1))
title('Subplot 1: Histogram PC1')

%Histograma para el PC2
subplot(2,2,2);
histogram(scorescod(:,2))
title('Subplot 2: Histogram PC2')

%Histograma para el PC3
subplot(2,2,[3,4]);
histogram(scorescod(:,3))
title('Subplot 3: Histogram PC3')

```



## CLUSTERIZACIÓN - KMEDOIDS

```
%Se observan las agrupaciones que se forman desde k=2 hasta k=7
k=7;
for i=2:k
    disp(["KMEDOIDS con K= ",num2str(i)])
    idxcmed=zeros(i,1);
    med=zeros(i,3);
    [idxmed,Cmed,sumdmed] =
kmedoids(scorescod(:,1:3),i,'Distance','sqeuclidean',"Algorithm","pam");
    figure
    scatter3(scorescod(:,1),scorescod(:,2),scorescod(:,3),20,idxmed)
    title('Cluster Assignments and Medoids');
    figure
    scatter(scorescod(:,1),scorescod(:,2),20,idxmed)
    hold on
    plot(Cmed(:,1),Cmed(:,2),'co',...
        'MarkerSize',7,'LineWidth',1.5)
    title('Cluster Assignments and Medoids PC1 vs PC2');
    hold off
    %     for j=1:i
    % %         idxcmed(j)=find(scorescod(:,1)==Cmed(j,1));
    % %         disp(["The medoid ", num2str(j)])
    % %         disp([CatVar(idxcmed(j),1), CatVar(idxcmed(j),2),
CatVar(idxcmed(j),3)])
    %     end
end

%Con las anteriores vistas, se encuentra que se forman grupos con
%muy pocos individuos, al revisar los histogramas realizados en la
%sección del PCA se identifican OUTLIERS

%Se remueven las zonas que están por encima del percentil 99
[scorescodWO, TF]=rmoutliers(scorescod(:,1:3),'percentiles',[0 99]);
DataWO=removerows(Matriz,"ind",TF(:,1)==1);

%Se observan como quedan los histogramas

%Histograma para el PC1
subplot(2,2,1);
histogram(scorescodWO(:,1))
title('Subplot 1: Histogram PC1 without Outliers')

%Histograma para el PC2
subplot(2,2,2);
```

```

histogram(scorescodWO(:,2))
title('Subplot 2: Histogram PC2 without Outliers')

%Histograma para el PC3
subplot(2,2,[3,4]);
histogram(scorescodWO(:,3))
title('Subplot 3: Histogram PC3 without Outliers')

%Se construye nuevamente la data para el PCA
PCAWO=[DataWO(:,9),DataWO(:,12:13),DataWO(:,15:18),DataWO(:,20:21)];
PCAWO=table2array(PCAWO);
datacentralizedWO=removerows(datacentralized,"ind",TF(:,1)==1);
CatVarWO=DataWO(:,1:8);

[coeffcentWO,scorecentWO,~,~,explainedcentWO,~] = pca(datacentralizedWO);
%Se observa que no se pierda información, la varianza explicada acumulada
%de las 3 primeras PCs no cambia significativamente

%Se evalua el número de cluster óptimo

eva =
evalclusters(scorescodWO(:,1:3),'kmeans','CalinskiHarabasz','KList',[1:7])%
5 --> 1.889e3 MAXIMIZE
eva2 =
evalclusters(scorescodWO(:,1:3),'kmeans','DaviesBouldin','KList',[1:7])%2 --
> 0.4845 MINIMIZE
%GAP
eva31 =
evalclusters(scorescodWO(:,1:3),'kmeans','gap','KList',[1:7],"Distance","sqE
uclidean") %5 -->1.6689
eva32 =
evalclusters(scorescodWO(:,1:3),'kmeans','gap','KList',[1:7],"Distance","cit
yblock") %7 -->0.9646
eva33 =
evalclusters(scorescodWO(:,1:3),'kmeans','gap','KList',[1:7],"Distance","cor
relation") %5 -->0.5325
eva34 =
evalclusters(scorescodWO(:,1:3),'kmeans','gap','KList',[1:7],"Distance","cos
ine") %7 -->0.2638
% "Euclidean" "Jaccard" "Hamming" NO SON ARGUMENTOS VALIDOS
%silhouette

eva41 =
evalclusters(scorescodWO(:,1:3),'kmeans','silhouette','KList',[1:7],"Distanc
e","sqEuclidean") %2 -->0.8406

```

```

eva42 =
evalclusters(scorescodWO(:,1:3),'kmeans','silhouette','KList',[1:7],"Distance",
"cityblock") %2 -->0.5862
eva43 =
evalclusters(scorescodWO(:,1:3),'kmeans','silhouette','KList',[1:7],"Distance",
"correlation") %4 -->0.8739
eva44 =
evalclusters(scorescodWO(:,1:3),'kmeans','silhouette','KList',[1:7],"Distance",
"cosine") %4 -->0.7166

%Por los índices, se descartan k=2, 6 y 7
%Se opta por escoger la configuración final de k=3, 4 y 5 mediante metodo
%gráfico

k=5;
for i=3:k
    disp(["KMEDOIDS con K= ",num2str(i)])
    idxcmedWO=zeros(i,1);
    medWO=zeros(i,3);
    [idxmedWO,CmedWO,sumdmedWO] =
kmedoids(scorescodWO(:,1:3),i,'Distance','sqeuclidean',"Algorithm","pam");
    figure
    scatter3(scorescodWO(:,1),scorescodWO(:,2),scorescodWO(:,3),20,idxmedWO)
    title('Cluster Assignments and Medoids');
    figure
    scatter(scorescodWO(:,1),scorescodWO(:,2),20,idxmedWO)
    hold on
    plot(CmedWO(:,1),CmedWO(:,2),'co',...
        'MarkerSize',7,'LineWidth',1.5)
    title('Cluster Assignments and Medoids PC1 vs PC2');
    hold off
    %     for j=1:i
    %         idxcmedWO(j)=find(scorescodWO(:,1)==CmedWO(j,1));
    %         disp(["The medoid ", num2str(j)])
    %         disp([CatVarWO(idxcmedWO(j),1), CatVarWO(idxcmedWO(j),2),
CatVarWO(idxcmedWO(j),3)])
    %     end
end

%Observando las vistas, se define k=5, como resultado óptimo.

```



## FICHA TÉCNICA FORMULARIOS / FORMATOS SUI

SERVICIO	ENERGÍA	CLASIFICACIÓN	N/A	FORMATO / FORMULARIO	FORMATO ZNI TO1
NOMBRE FORMATO	REGISTRO DE OPERACIÓN DIARIO			PERIODICIDAD REPORTE	TRIMESTRAL - SEMESTRAL
TÓPICO AÑOS DISPONIBLES	TÉCNICO 2003-2017	ACTO ADMINISTRATIVO	CIRCULAR SSPD - IPSE NO. 0001 DE 2004	No. CAMPOS	17
DEFINICION DE CAMPOS					
<b>Identificador_Empresa</b>	Corresponde al código de identificación en base de datos, de la empresa que realiza el reporte de la información al SUI.				
<b>Empresa</b>	Nombre de la empresa que reporta la información, conforme a como esta registrado en el Registro Único de Prestadores de Servicios - RUPS.				
<b>Código DANE</b>	Corresponde a la codificación dada por el DANE a la división político-administrativa de Colombia. Con la siguiente estructura: DDMMMCCC, donde "DD" es el código del departamento, "MMM" corresponde al código del municipio y "CCC" corresponde al código del centro poblado. Para los casos en que no aplique el centro poblacional se debe diligenciar 000.				
<b>Departamento</b>	Corresponde al Departamento donde está se realiza la prestación del servicio				
<b>Municipio</b>	Corresponde al Municipio donde está se realiza la prestación del servicio				
<b>Centro Poblado</b>	Corresponde al Centro Poblado donde está se realiza la prestación del servicio				
<b>Nombre Localidad</b>	Corresponde a la numeración que se tiene de la planta.				
<b>Localidad</b>	Hace referencia al nombre de la cabecera o localidad menor que se está reportando.				
<b>Planta Eléctrica</b>	Corresponde a la numeración que se tiene de la planta.				
<b>Marca</b>	Corresponde a la marca del grupo electrógeno que opere a la hora indicada en la columna 6.				
<b>Fecha</b>	Se refieren a la fecha en que fue prestado el servicio y de la cual se están reportando los datos de operación del grupo electrógeno.				
<b>Horas de Servicio</b>	Corresponde a las horas de prestación de la planta en la fecha indicada en el punto 5.				
<b>Capacidad de Generación</b>	Corresponde a la capacidad del grupo electrógeno que opere a la hora indicada en la columna 6.				
<b>Horómetro</b>	Corresponde a las horas acumuladas de funcionamiento al momento de la lectura a la hora señalada en el campo 6.				
<b>kW Generados</b>	Corresponde la cantidad de kW generados cada hora; este puede ser tomado del tablero de medida en el momento de la lectura de la hora señalada en el campo 6 o puede ser calculado de forma indirecta mediante las variables de corriente, tensión y factor de potencia.				
<b>Año</b>	Año de cargue de la información				
<b>Periodo</b>	Corresponde al Periodo de cargue de la información				
				<b>FECHA ACTUALIZACIÓN FICHA TÉCNICA</b>	1/12/2017

 <b>Superservicios</b> Superintendencia de Servicios Públicos Domiciliarios		<b>FICHA TÉCNICA FORMULARIOS / FORMATOS SUI</b>			
<b>SERVICIO</b>	ENERGIA	<b>CLASIFICACION</b>	N/A	<b>FORMATO / FORMULARIO</b>	FORMATO ZNI C5
<b>NOMBRE FORMATO</b>	REPORTE USO DE LOS SUBSIDIOS POR MENORES TARIFAS			<b>PERIODICIDAD REPORTE</b>	TRIMESTRAL - SEMESTRAL
<b>TÓPICO AÑOS DISPONIBLES</b>	COMERCIAL 2014-2017	<b>ACTO ADMINISTRATIVO</b>	CIRCULAR SSPD - IPSE NO. 0001 DE 2004	<b>No. CAMPOS</b>	19
<b>DEFINICION DE CAMPOS</b>					
<b>Identificador Empresa</b>	Corresponde al código de identificación en base de datos, de la empresa que realiza el reporte de la información al SUI.				
<b>Empresa</b>	Nombre de la empresa que reporta la información, conforme a como esta registrado en el Registro Único de Prestadores de Servicios - RUPS.				
<b>Año</b>	Corresponde al año de reporte de la información.				
<b>Trimestre</b>	Corresponde al trimestre de reporte de la información.				
<b>Cód. DANE</b>	Corresponde a la codificación dada por el DANE a la división político-administrativa de Colombia. Con la siguiente estructura: DDMMMCCC, donde DD es el código del departamento, MMM corresponde al código del municipio y CCC corresponde al código del centro poblado. Para los casos en que no aplique el centro poblacional se debe diligenciar 000.				
<b>Centro Poblado</b>	Corresponde al nombre de la cabecera municipal y localidad menor en mayúsculas como aparece en las resoluciones del IPSE.				
<b>Fecha</b>	Se refiere a la fecha en la cual se realizó la transacción consignada en el comprobante de egreso, en el caso del reporte de los montos subsidiados a los estratos 1, 2 y 3, se debe corresponden a la fecha en que se emitió la facturación.				
<b>Comprobante Egreso</b>	Corresponde al serial asignado al comprobante de egreso; aplica para todos gastos excepto para los montos subsidiados a los estratos 1, 2 y 3.				
<b>Galones Comprados</b>	Hace referencia a la cantidad de galones de combustibles comprados.				
<b>Valor Combustible</b>	Corresponde al valor en pesos por cada galón de combustible.				
<b>Valor Total</b>	Corresponde al producto del valor del combustible por galón, multiplicado por la cantidad de galones de combustible comprados.				
<b>Valor Mantenimiento G Electrónico</b>	Hace referencia al valor pagado por el mantenimiento eléctrico, mecánico, electrónico y demás, realizado en el Grupo electrógeno.				
<b>Valor Mantenimiento Redes</b>	Hace referencia al valor pagado por el mantenimiento a las redes de transmisión y/o distribución.				
<b>Subsidios Facturación Estrato1</b>	Corresponde al valor de subsidio facturado a los usuarios de estrato 1.				
<b>Subsidios Facturación Estrato2</b>	Corresponde al valor de subsidio facturado a los usuarios de estrato 2.				
<b>Subsidios Facturación Estrato3</b>	Corresponde al valor de subsidio facturado a los usuarios de estrato 3				
<b>Otros Egresos</b>	Corresponde a otros egresos no tenidos en cuenta en las variables anteriores y los cuales se relacionan con el uso de subsidios por menores tarifas en cuanto a compra de energía para agentes que destinan este recurso para el pago de energía al agente generador cuando el prestador no presta esta actividad y destina este recurso para el pago de la energía.				
<b>Valor Transporte Combustible</b>	Este campo corresponde al valor del transporte a cada cabecera o centro poblado (terrestre, fluvial y marítimo).				
<b>Punto Acopio Combustible</b>	Se refiere al punto de entrega del combustible donde el prestador recibe el combustible.				
				<b>FECHA ACTUALIZACION FICHA TÉCNICA</b>	30/11/2017

ANEXO: Muestra de la data de biomasa

Departamento	Municipio	Localidad	Factor de ponderación del PE por usuarios	Potencial anual bruto arroz kWh/año	Potencial anual bruto maíz kWh/año	Potencial anual bruto caña kWh/año	Potencial anual bruto palma kWh/año	Potencial anual bruto banano kWh/año	Potencial anual bruto plátano kWh/año	Potencial bruto biomasa kWh/año	Potencial bruto biomasa seca con FR	Potencial bruto biomasa húmeda con FR	Potencial técnico biomasa seca	Potencial técnico biomasa húmeda	Energía anual diseño [kWh]	Margen con biomasa seca	Margen con biomasa húmeda	Margen total
CHOCO	RIOSUCIO	caño seco	0.080	25841080	2594177	40961	6457	414304	20073436	48970415	22787431	16390192	2962366	2130725	30240	97.96	70.46	168.42
ANTIOQUIA	TURBO	Bocas de atrato	1.000	46874769	11827223	0	0	738046446	774951984	1571700422	46961594	1210398744	6105007	157351837	73440	83.13	2142.59	2275.72
CHOCO	RIOSUCIO	dos bocas	0.064	20429357	2050896	32383	5104	327539	15869592	38714870	18015213	12957705	2341978	1684502	36180	64.73	46.56	111.29
VICHADA	LA PRIMAVERA	santa barbara	1.000	181034	1505284	932660	5750277	0	993519	9362775	7845460	794815	1019910	103326	35100	29.06	2.94	32.00
CHOCO	NOVITA	santa barbara	1.000	4215187	3206078	2307079	0	4600215	2822599	17151159	7782676	5938251	1011748	771973	35100	28.82	21.99	50.82
CHOCO	RIOSUCIO	la nueva	0.015	4714039	473241	7472	1178	75579	3661881	8933389	4156979	2989968	540407	388696	18950	28.52	20.51	49.03
CHOCO	RIOSUCIO	guineo	0.007	2306870	231586	3657	576	36986	1791984	4371659	2034266	1463176	264455	190213	9274	28.52	20.51	49.03
CHOCO	CARMEN DEL DARIEN	manuel santo	0.102	1978213	713800	0	0	222948	24173820	27088781	2153611	19517414	279969	2537264	11693	23.94	216.99	240.94
CHOCO	RIOSUCIO	regadero	0.027	8826285	886068	13991	2205	141510	6856288	16726346	7783280	5598238	1011826	727771	47520	21.29	15.32	36.61
CHOCO	RIOSUCIO	salaquisito	0.026	8425091	845792	13355	2105	135077	6544638	15966058	7429495	5343772	965834	694690	45360	21.29	15.32	36.61
CHOCO	RIOSUCIO	jhuinpubur indigena	0.032	10330766	1037102	16375	2581	165631	8024973	19577428	9109976	6552483	1184297	851823	55620	21.29	15.32	36.61
CHOCO	RIOSUCIO	clevellino	0.023	7321805	735033	11606	1829	117389	5687602	13875264	6456585	4643993	839356	603719	39420	21.29	15.32	36.61
CHOCO	CARMEN DEL DARIEN	chintado medio	0.170	3297022	1189667	0	0	371580	40289700	45147969	3589351	32529024	466616	4228773	33480	13.94	126.31	140.24
CHOCO	MEDIO BAUDO	playa alta	0.020	616126	2126	85661	0	0	731267	1435181	563131	585014	73207	76052	5645	12.97	13.47	26.44

ANEXO: Muestra de la data empleada para el PCA

Departamento	Municipio	Localidad	Capacidad [kW]	Usuarios E1	Energía Facturada E1 [kWh/mes]	Costo unitario [\$/KWh]	Factura E1 [\$/mes]	Subsidio total	Diesel comprado [Gal]	Precio por galon [\$/Gal]	Costo transporte diesel [\$/Gal]
CHOCO	RIOSUCIO	caño seco	40.000	56	2264	1480	3351449	2737514	245	6312	335421
ANTIOQUIA	TURBO	Bocas de atrato	67.000	136	9126	1162	10394797	7823039	441	6940	597108
CHOCO	RIOSUCIO	dos bocas	60.000	67	2719	1480	4025366	3287979	294	6312	402505
VICHADA	LA PRIMAVERA	santa barbara	115.000	68	2834	1159	3247594	2490016	600	6703	3300
CHOCO	NOVITA	santa barbara	65.000	65	2633	1493	3930875	3216785	285	6674	390199
CHOCO	RIOSUCIO	la nueva	60.000	47	1416	1480	2096457	1712417	153	6312	209467
CHOCO	RIOSUCIO	guineo	20.000	23	699	1480	1034007	844592	76	6312	104049
CHOCO	CARMEN DEL DARIEN	manuel santo	19.000	29	1404	1328	1863936	1070944	137	14000	2500
CHOCO	RIOSUCIO	regadero	90.000	88	3563	1480	5274994	4308693	386	6312	528459
CHOCO	RIOSUCIO	salaquisito	65.000	84	3407	1480	5044278	4120241	369	6312	505185
CHOCO	RIOSUCIO	jhuinpubur indigena	67.200	103	4155	1480	6150435	5023766	450	6312	616079
CHOCO	RIOSUCIO	clevellino	60.000	73	2962	1480	4385151	3581856	321	6312	439470
CHOCO	CARMEN DEL DARIEN	chintado medio	19.000	62	3949	1328	5242321	3134576	386	14000	2500
CHOCO	MEDIO BAUDO	playa alta	12.000	14	462	1738	802799	677509	46	6668	29468

ANEXO: Muestra demostrativa del cálculo del indicador compuesto

Departamento	Municipio	Localidad	nDAU	nDD	nTBM	Alpha	Betha	Delta	Gamma	AporteDAUyDD	AporteBM	nCTTB	%ApDAUyDD	%ApBM	Prod%Ap	FC	nCTBcorr
CHOCO	RIOSUCIO	caño seco	0.626	0.491	1.000	0.104	0.113	0.097	1.000	0.084	0.500	0.584	0.144	0.856	0.123	0.493	0.288
ANTIOQUIA	TURBO	Bocas de atrato	0.337	0.623	1.000	0.458	0.182	0.105	1.000	0.167	0.500	0.667	0.250	0.750	0.188	0.750	0.500
CHOCO	RIOSUCIO	dos bocas	0.624	0.526	1.000	0.128	0.113	0.076	1.000	0.090	0.500	0.590	0.152	0.848	0.129	0.515	0.304
VICHADA	LA PRIMAVERA	santa barbara	0.649	0.416	0.991	0.182	0.003	0.082	1.000	0.077	0.495	0.572	0.134	0.866	0.116	0.464	0.266
CHOCO	NOVITA	santa barbara	0.625	0.542	1.000	0.123	0.154	0.067	1.000	0.098	0.500	0.598	0.164	0.836	0.137	0.549	0.329
CHOCO	RIOSUCIO	la nueva	0.627	0.582	1.000	0.061	0.113	0.036	1.000	0.062	0.500	0.562	0.111	0.889	0.098	0.394	0.221
CHOCO	RIOSUCIO	guineo	0.623	0.530	1.000	0.024	0.113	0.057	1.000	0.052	0.500	0.552	0.095	0.905	0.086	0.343	0.190
CHOCO	RIOSUCIO	regadero	0.625	0.544	1.000	0.171	0.113	0.066	1.000	0.102	0.500	0.602	0.170	0.830	0.141	0.563	0.339
CHOCO	RIOSUCIO	salaquisito	0.625	0.503	1.000	0.163	0.113	0.089	1.000	0.102	0.500	0.602	0.169	0.831	0.141	0.562	0.338
CHOCO	RIOSUCIO	jhuinpubur indigena	0.627	0.475	1.000	0.202	0.113	0.107	1.000	0.115	0.500	0.615	0.188	0.812	0.152	0.609	0.375
CHOCO	RIOSUCIO	clevellino	0.624	0.512	1.000	0.140	0.113	0.084	1.000	0.094	0.500	0.594	0.158	0.842	0.133	0.533	0.317
CHOCO	MEDIO BAUDO	playa alta	0.585	0.485	1.000	0.011	0.071	0.058	1.000	0.034	0.500	0.534	0.064	0.936	0.060	0.241	0.129