Proceso de Descubrimiento de Reglas de Caracterización de Grupos Espacialmente Referenciados

Giovanni Daián Rottoli[1,2,3], Hernán Merlino[3], Ramón García-Martinez[3]

[1] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. Argentina.
[2] Programa de Becas "Formación de Doctores para Fortalecer áreas de I+D+i". Universidad Tecnológica Nacional. Argentina.
[3] Grupo de Investigación en Sistemas de Información. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina.
gd.rottoli@gmail.com, hmerlino@gmail.com, rgm1960@yahoo.com

Resumen. El descubrimiento de grupos sobre información espacialmente referenciada es un tema de interés en minería de datos espaciales. Los algoritmos existentes se orientan a descubrir tres tipos de grupos: regiones, grupos propiamente dichos, y zonas calientes. Sin embargo, no dan una caracterización de dichos grupos. El presente trabajo propone un proceso de descubrimiento de reglas de caracterización de grupos espacialmente referenciados que utiliza algoritmos TDIDT para obtener las características que fueron elegidas automáticamente para la generación de los mismos. Se presentan pruebas de concepto en las que se utilizan distintos algoritmos de agrupamiento espacial.

Palabras Clave. Reglas de Caracterización de Grupos, Agrupamiento, Información Espacial, TDIDT, Explotación de Información.

1 Introducción

El descubrimiento de grupos, o *clustering* por su término en inglés, permite identificar particiones de la masa de información sobre un dominio de problema determinado [Britos, 2008]. La incorporación de información espacialmente referenciada a esta tarea agrega dificultad a la misma [Kataria & Rupal, 2012], haciendo que nuevos algoritmos y métodos nuevos hayan surgido en los últimos años. Las distintas formas de considerar esta información espacial y las restricciones tenidas en cuenta para la generación de grupos permiten diferenciar tres tipos de abordajes diferentes: regionalización, descubrimiento de zonas calientes o *hotspots* y descubrimiento de grupos espaciales propiamente dicho, concepto que utilizaremos para referirnos a los tres enfoques en forma conjunta [Mennis & Guo, 2009].

En principio, la generación de grupos espaciales es el caso más general, en el cual se buscan conjuntos disjuntos utilizando atributos espaciales, no espaciales, o ambos [Liu et al., 2012]. Por otro lado, la regionalización agrega al descubrimiento de grupos una condición de contigüidad entre los objetos espaciales, permitiendo encontrar conjuntos de objetos contiguos cuyos valores de sus atributos permitan

optimizar cierta función objetivo, tratándose usualmente de una función de homogeneidad o heterogeneidad [Guo, 2008]. Por último, la utilización de los algoritmos de agrupamiento espacial se aplica al descubrimiento de zonas calientes o hotspots, zonas del espacio en las cuales existen concentraciones de puntos espaciales inusuales, haciéndose uso de algoritmos de clustering basados en densidades [Brimicombe, 2007; Nisa et al., 2014; Santoso & Nisa, 2016].

La creación y mejora de algoritmos para estos objetivos ha sido, como se mencionó anteriormente, una tarea constante en los últimos años, pudiéndose encontrar ejemplos de estos en [Brimicombe, 2007; Guo, 2008; Yang & Cui, 2008; Mennis & Guo, 2009; Zhong et al., 2010; Deng et al., 2011; Shah et al., 2012, Popat & Emmanuel, 2014]. Sin embargo, ninguno de estas propuestas brinda en sí misma la posibilidad de caracterizar los grupos generados automáticamente, de forma tal de poder conocer cuáles fueron los criterios utilizados por los algoritmos para realizar esta actividad.

En el presente trabajo, y con base en los hechos presentados hasta el momento, se presenta en la sección 2 la problemática derivada, en la sección 3 una solución propuesta a la misma, pruebas de concepto de dicha solución en la sección 4, y por último se presentan conclusiones en la sección 5.

2 Definición del Problema

Muchos son los algoritmos y métodos desarrollados para el descubrimiento de grupos espaciales en cualquiera de sus formas: grupos, regiones o zonas calientes, tanto para puntos en el espacio como para otros tipos de datos espaciales. Sin embargo, como se mencionó en el apartado anterior, los algoritmos y métodos desarrollados no permiten caracterizar los grupos espaciales descubiertos, en función de los atributos elegidos para tal actividad.

Por similitud con [Britos, 2008], resulta interesante la creación de un proceso para la caracterización de grupos espacialmente referenciados que permita el descubrimiento de grupos y su descripción posterior mediante reglas, de manera sistemática, independientemente del abordaje que se seleccione para la generación de los mismos. Por esta razón, se plantea un proceso de explotación de información para la obtención de reglas de caracterización de grupos espacialmente referenciados utilizando algoritmos TDIDT.

3 Solución Propuesta

[Britos, 2008] propone un proceso de explotación de información para el descubrimiento de reglas de pertenencia a grupos, en el cual utiliza algoritmos TDIDT sobre el resultado de un proceso de descubrimiento de grupos para hallar las reglas que caracterizan a cada uno de los grupos generados. Bajo el mismo concepto, se propone un proceso de explotación de información denominado Proceso de Descubrimiento de Reglas de Caracterización de Grupos Espacialmente Referenciados, que permite determinar las características de los objetos agrupados en un mismo grupo espacial, independientemente de si son de regiones, grupos

espaciales o zonas calientes. Para tal fin, tal como se puede observar en la Figura 1, a partir de un conjunto de información espacialmente referenciada disponible en distintas fuentes y formatos, se realiza un proceso de integración para obtener otro conjunto de información espacial integrada que conste de atributos espaciales, tales como la ubicación espacial del objeto, y atributos no espaciales, en un único registro.

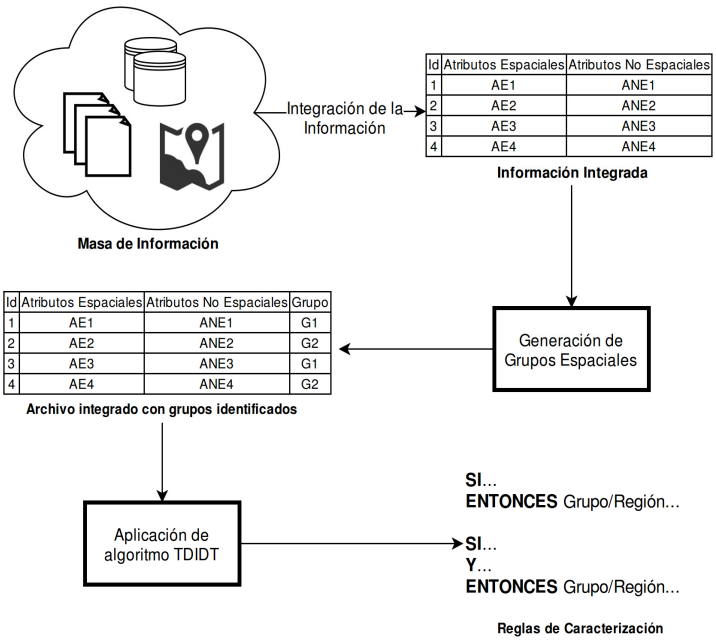

Fig. 1. Proceso de descubrimiento de reglas de caracterización de grupos espacialmente referenciados, donde la sigla AE significa Atributo Espacial y la sigla ANE Atributo No-Espacial

La información integrada es sometida posteriormente a un proceso de descubrimiento de grupos, como se puede observar en la Figura 2. En este paso es necesario optar por una de las tres clases de grupos espaciales mencionadas anteriormente en función al dominio del problema que se esté abordando y seleccionar los algoritmos adecuados en cada oportunidad: para la generación de regiones se sugiere la utilización de algoritmos de la familia REDCAP, debido a sus ventajas sobre otros algoritmos [Guo, 2008]. En este caso, es necesario, además, identificar la relación de contigüidad entre los objetos espaciales. Por otro lado, tanto en el caso del descubrimiento de grupos espaciales como de zonas calientes, se sugiere la utilización de algoritmos basados en densidad tales como aquellos de la familia DBSCAN [Ester et al., 1996; Sander et al., 1998; Nisa et al., 2014], DENCLUE [Hinneburg & Keim, 1998], ASCDT [Deng et al., 2011] o DBSC [Liu et al., 2012], por los mismos motivos mencionados anteriormente [Shah et al., 2012; Popat & Emmanuel, 2014]. En cada caso, del algoritmo seleccionado dependen los atributos de entrada que se le proporcionará. Como resultado de esta etapa se obtiene

un archivo en el cual consta la información integrada más un nuevo atributo en el cual se especifica el grupo espacial al que pertenece.

En un paso posterior, el archivo generado es utilizado como entrada de un algoritmo *Top-Down Induction of Decision Trees* – TDIDT – para la generación de las reglas de caracterización de grupos espacialmente referenciados, especificando el atributo agregado en el paso anterior, "Grupo", como atributo objetivo o *target*, y los atributos no espaciales como atributos de entrada, tal como se puede observar en la Figura 3.

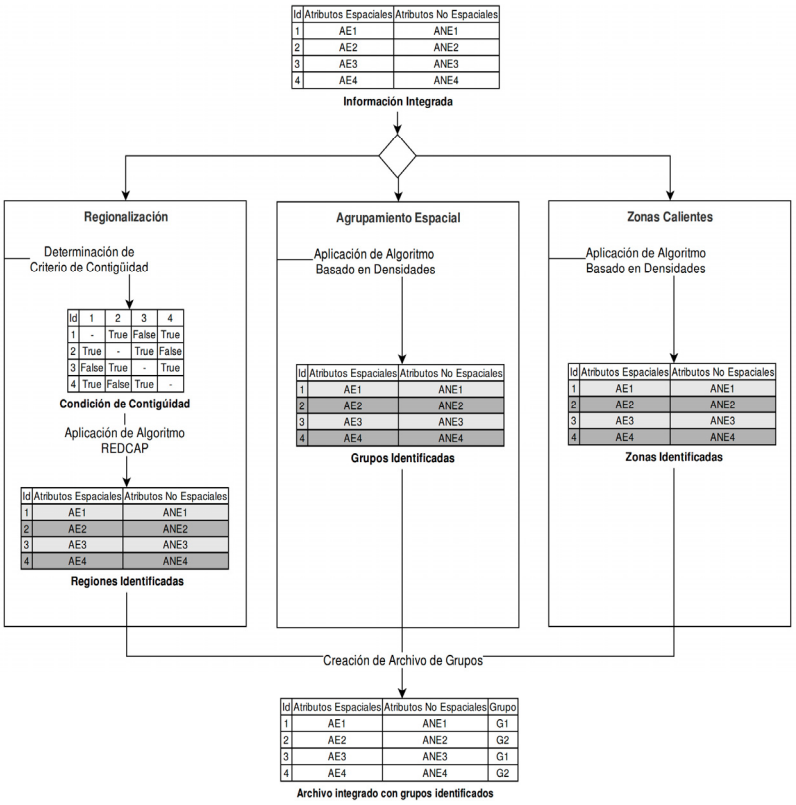

Fig. 2. Subproceso para agrupación de datos espaciales, donde la sigla AE significa Atributo Espacial y la sigla ANE significa Atributo No-Espacial

4 Pruebas de Concepto

En esta sección se presentan dos pruebas de concepto del proceso propuesto utilizando datos reales obtenidos de distintas fuentes. En la sección 4.1 se utiliza algoritmos de regionalización sobre datos referidos a las provincias de Argentina, mientras que en la sección 4.2 se hace uso de algoritmos de descubrimiento de grupos espaciales sobre datos recogidos diariamente por distintas estaciones meteorológicas de Argentina.

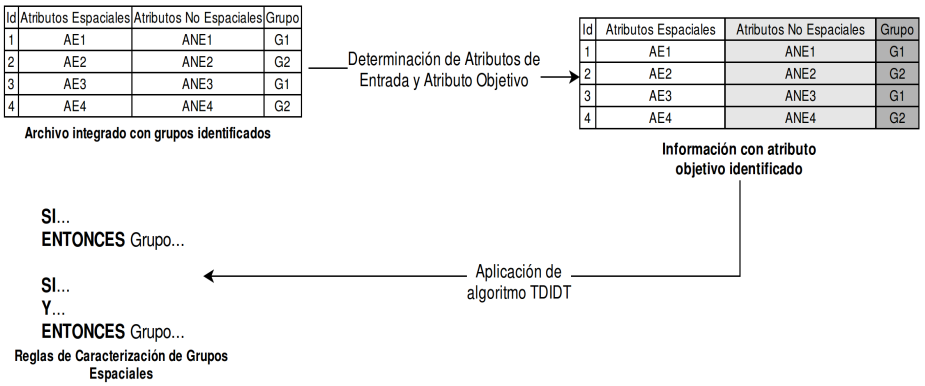

Fig. 3. Subproceso para obtención de reglas de caracterización, mediante la aplicación de Algoritmos TDIDT

4.1 Regionalización de provincias Argentinas

Se realiza la primera prueba de concepto utilizando datos obtenidos en el Censo Nacional de Población, Hogares y Viviendas realizado por la Institución Nacional de Estadísticas y Censos de Argentina en el año 2010 [INDEC, 2010a, 2010b, 2010c, 2010d, 2010e].
Los datos en cuestión han sido integrados y normalizados en un conjunto de atributos, tal como se muestra en la Tabla 1, siendo estos de tipo numérico a excepción del atributo Provincia, el cual corresponde a un valor alfanumérico.

Tabla 1. Descripción de atributos de los datos de prueba sobre las provincias argentinas

Atributo	Descripción
Provincia	Nombre de la provincia
TotViv	Total de viviendas en cada provincia
Superf	Superficie de la provincia, en Km^2
Pob	Cantidad de Habitantes
Dens	Cantidad de personas por kilómetro cuadrado
Analf	Cantidad de personas analfabetas por provincia
EdadM	Edad mediana del total de la población
Porc +65	Porcentaje de la población mayor a 65 años respecto al total de la población
PobOrig	Cantidad de personas descendientes de pueblos originarios con viviendas particulares

Los datos integrados son utilizados como entrada de un algoritmo de regionalización REDCAP. Para este experimento se ha seleccionado el algoritmo First Order SLK [Guo, 2008] debido a la simplicidad de su implementación, usando además, como criterio de contigüidad, la vecindad real de las provincias argentinas, generándose seis regiones espaciales tal como se puede observar en la Figura 5.

Posteriormente, se integran las regiones identificadas en un único archivo de datos en el que constan tanto los datos de entrada originales como la región a la cual pertenece cada uno de ellos. Dicho archivo se utiliza como entrada del algoritmo TDIDT Random Tree [Breiman, 2001] utilizando Tanagra [Rakotomalala, 2005] para su ejecución, obteniéndose un árbol de decisión que se deriva en las reglas que se pueden observar en la Figura 4, de la cual se puede extraer las características de cada grupo destacándose que la confianza del obtenida es del 100%.

SI PobOrig < 49703,5
Y Analf >= 14444,5
Y Superf >= 26162,5
Y Superf < 139895,5
Y Dens < 10,9227
ENTONCES Región 1

SI PobOrig >= 49703,5
Y TotViv < 775948,5
ENTONCES Región 1

SI PobOrig < 49703,5
Y Analf >= 14444,5
Y Superf >= 26162,5
Y Superf < 139895,5
Y Dens >=10,9227
ENTONCES Región 2

SI PobOrig < 49703,5
Y Superf < 139895,5
Y Analf < 14444,5
Y Porc +65 >= 6,86%
ENTONCES Región 3

SI PobOrig < 49703,5
Y Analf >= 14444,5
Y Superf < 26162,5
ENTONCES Región 3

SI PobOrig >= 49703,5
Y TotViv >= 775948,5
ENTONCES Región 4

SI PobOrig < 49703,5
Y Superf < 139895,5
Y Analf < 14444,5
Y Porc +65 < 6,86%
ENTONCES Región 5

SI PobOrig < 49703,5
Y Superf >= 139895,5
Y EdadM >= 28,95
ENTONCES Región 5

SI PobOrig < 49703,5
Y Superf >= 139895,5
Y EdadM < 28,95
ENTONCES Región 6

Fig. 4. Reglas de Caracterización de grupos espacialmente referenciados de la primera prueba de concepto

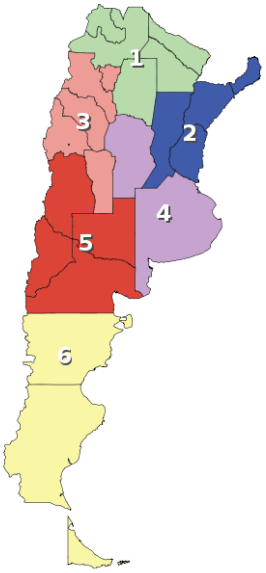

Fig. 5. Mapa de Argentina con regiones generadas

Resulta interesante analizar las características de las regiones descubiertas. En primera instancia se puede observar que el principal atributo para la decisión es la cantidad de habitantes de pueblos originarios: aquellas regiones con mayor cantidad son la región 1, correspondiente al noroeste argentino, y la región 4, correspondiente a Buenos Aires, Córdoba y la Ciudad Autónoma de Buenos Aires, teniendo esta última región mayor cantidad de viviendas que la primera. Por otro lado, aquellas regiones con menor cantidad de habitantes originarios se caracterizan por su superficie, cantidad de personas analfabetas, el porcentaje de personas mayores a 65 años, y la densidad poblacional.

La región 5 y la región 6 se caracterizan por poseer una mayor extensión territorial, y se distinguen entre sí según la edad mediana de la población, teniendo la primera mayor edad mediana que la segunda, siendo el límite aproximadamente 29 años de edad. Cabe destacar que, según las reglas encontradas, existe una región con superficie menor que las restantes pertenecientes al mismo grupo, similar en características a la región 3, la cual posee en su mayoría baja tasa de analfabetismo.

Aquellas regiones con superficie pequeña y mayor tasa de analfabetismo se diferencian a su vez según su densidad poblacional. Aquellas con una densidad menor a 11 personas por Km^2 corresponden a algunas provincias de grupo 1, y si por el contrario, la densidad es mayor a 11 personas por Km^2, a algunas provincias del grupo 2.

4.2 Agrupamiento de estaciones meteorológicas argentinas

Se utilizaron datos recolectados por estaciones meteorológicas de Argentina el día 6 de junio de 2016 [INTA, 2016]. Se utilizaron los atributos que se pueden observar en la Tabla 2, utilizando para el agrupamiento basado en densidades utilizando DBSCAN [Ester et al., 1996] mediante el software WEKA [Hall et al., 2009], indicando como parámetros *eps* y *minpts* del algoritmo los valores 0,075 y 4 respectivamente, resultando 5 grupos espaciales, con la distribución que se observa en la Figura 7. Para este paso, los valores fueron normalizados.

Posteriormente, dichos grupos espaciales son integrados al archivo de datos y utilizados como entrada del algoritmo C4.5 [Quinlan, 1993] mediante el software Tanagra, utilizando como atributo objetivo el grupo descubierto y como atributos de entrada las temperaturas mínima, máxima y media, con valores sin normalizar, resultando un árbol de decisión que se deriva en las reglas que se pueden observar en la Figura 6.

Tabla 2. Descripción de atributos de los datos de centrales meteorológicas argentinas considerados en la prueba de concepto

Atributo	Descripción
Lat	Latitud de las coordenadas de la central meteorológica.
Long	Longitud de las coordenadas de la central meteorológica.
TMin	Temperatura Mínima medida en el día.
TMax	Temperatura Máxima medida en el día.
TMed	Temperatura Media del día. (Dato calculado)

```
SI TMed < 9,85                SI TMed >= 9,85               SI TMed >= 9,85
Y TMin > -0,65                Y TMin >= 7,2                 Y TMin < 7,2
Y TMax < 13,5                 ENTONCES Grupo 2              ENTONCES Grupo 4
ENTONCES Grupo 0              SI TMed < 9,85
SI TMed < 9,85                Y TMin > -0,65
Y TMin < -0,65                Y TMax >= 13,5
ENTONCES Grupo 1              ENTONCES Grupo 3
```

Fig. 6. Reglas de caracterización de grupos espacialmente referenciados obtenidas como resultado de la aplicación de TDIDT sobre los datos de centrales meteorológicas argentinas

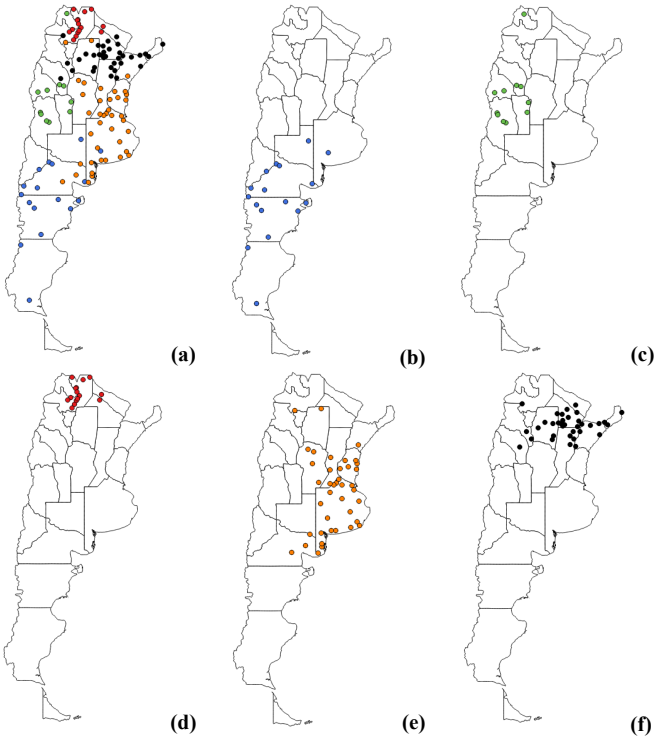

Fig. 7. (a) Distribución de todas las estaciones meteorológicas de Argentina con datos el día 6 de junio de 2016. (b) Estaciones meteorológicas en el grupo 0. (c) Estaciones meteorológicas en el grupo 1. (d) Estaciones meteorológicas en el grupo 2. (e) Estaciones meteorológicas en el grupo 3. (f) Estaciones meteorológicas en el grupo 4.

Las reglas de caracterización permiten diferenciar los distintos grupos con una certeza del 83,19%. En primer lugar, el grupo 0 posee una temperatura media menor a 9,85ºC, temperatura mínima mayor a -0,65ºC y una temperatura máxima menor a 13,5ºC, distinguiéndose por este último valor del grupo 3, el cual posee una temperatura máxima mayor o igual a 13,5ºC.

Por otro lado, el grupo 1 se caracteriza por sus bajas temperaturas media y mínima, siendo estas menores a 9,85ºC y -0,65ºC respectivamente, y siendo estos mismos valores mayores a 9,85ºC y 7,2ºC en el grupo 2.

Por último, el grupo 4 se caracteriza por temperaturas medias mayores a 9,85ºC y temperaturas mínimas menores a 7,2ºC.

5 Conclusiones

Se ha diseñado un proceso de explotación de información para el descubrimiento de reglas de caracterización de grupos espacialmente referenciados que permite la generación automática de grupos de objetos espaciales y la obtención de las características de cada uno de ellos utilizando algoritmos TDIDT. La caracterización

de estos grupos independientemente de que sean regiones, puntos calientes o grupos propiamente dichos.

Se han presentado dos pruebas de concepto que ilustran el funcionamiento del proceso propuesto utilizando datos reales. La primera utiliza algoritmos de generación de regiones, y la segunda utiliza algoritmos de generación de grupos. En ambos casos el algoritmo utilizado se compuso con un algoritmo de la familia TDIDT.

Queda pendiente evaluar el proceso propuesto para la caracterización de zonas calientes, que en esta etapa no se ha hecho por estar involucrada la misma familia de algoritmos que en la caracterización de grupos espaciales propiamente dicho.

Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Programa Formación de Doctores para Fortalecer áreas de I+D+i (2016-2020) de la Universidad Tecnológica Nacional (Argentina) y por los Proyectos de Investigación 33B133 y 33A205 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

Referencias

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brimicombe, A. J. (2007). A dual approach to cluster discovery in point event data sets. Computers, environment and urban systems, 31(1), 4-18.

Britos, P. V. (2008). Procesos de explotación de información basados en sistemas inteligentes Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.

Deng, M., Liu, Q., Cheng, T., & Shi, Y. (2011). An adaptive spatial clustering algorithm based on Delaunay triangulation. Computers, Environment and Urban Systems, 35(4), 320-332.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science, 22(7), 801–823.

Hinneburg, A., & Keim, D. A. (1998, August). An efficient approach to clustering in large multimedia databases with noise. In KDD (Vol. 98, pp. 58-65).

INDEC. (2010a). Cuadro V1. Total del país. Total de viviendas por provincia. Año 2010. Disponible en: ww.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=1&arch=x&c=2010. Accedido el 1 de junio de 2016.

INDEC. (2010b). Cuadro P 3. Total del país. Población total, superficie y densidad por provincia. Años 2001~2010. Disponible en: www.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=10&arch=x&c=2010. Accedido el 1 de junio de 2016.

INDEC. (2010c). Cuadro P17. Total del país. Edad mediana de la población por sexo, según provincia. Año 2010.Disponible en: www.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=24&arch=x&c=2010. Accedido el 1 de junio de 2016.

INDEC. (2010d). Cuadro P18. Total del país. Envejecimiento de la población por provincia, según censos nacionales 1970 a 2010. Disponible en: www.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=25&arch=x&c=2010. Accedido el 1 de junio de 2016.

INDEC. (2010e). Cuadro P46. Total del país. Población indígena o descendiente de pueblos indígenas u originarios en viviendas particulares por tipo de cobertura de salud, según provincia. Año 2010. Disponible en: www.indec.gov.ar/definitivos_bajarArchivoNacionales.asp?idc=62&arch=x&c=2010. Accedido el 1 de junio de 2016.

INTA. (2016). Datos Diarios. SIGA - Sistema de Información y Gestión Agrometeorológic. Instituto Nacional de Tecnología Agropecuaria. Argentina. Disponible en: http://siga2.inta.gov.ar/en/datosdiarios/. Accedido el 6 de junio de 2016

Kataria, P., & Rupal, N.(2012). Mining Spatial Data & Enhancing Classification Using Bio-Inspired Approaches. International Journal of Science and Research (IJSR). 3(2), 1473 – 1479.

Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers & Geosciences, 46, 296-309.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

Nisa, K. K., Andrianto, H. A., & Mardhiyyah, R. (2014, October). Hotspot clustering using DBSCAN algorithm and shiny web framework. In Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on (pp. 129-132). IEEE.

Popat, S. K., & Emmanuel, M. (2014). Review and Comparative Study of Clustering Techniques. International Journal of Computer Science and Information Technologies, 5(1), 805-812.

Quinlan, J. R. (1993). C4. 5: programs for machine learning.

Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. In Proceedings of EGC (Vol. 2, pp. 697-702).

Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.

Santoso, A., & Nisa, K. K. (2016, January). Cloud Computing Application for Hotspot Clustering Using Recursive Density Based Clustering (RDBC). In IOP Conference Series: Earth and Environmental Science (Vol. 31, No. 1, p. 012004). IOP Publishing.

Shah, G. H., Bhensdadia, C. K., & Ganatra, A. P. (2012). An empirical evaluation of density-based clustering techniques. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 2231-2307.

Yang, X., & Cui, W. (2008, December). A novel spatial clustering algorithm based on Delaunay triangulation. In International Conference on Earth Observation Data Processing and Analysis (pp. 728530-728530). International Society for Optics and Photonics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Zhong, C., Miao, D., & Wang, R. (2010). A graph-theoretical clustering method based on two rounds of minimum spanning trees. Pattern Recognition, 43(3), 752-766.