Información Tecnológica
Vol. 25(5), 95-102 (2014)
doi: 10.4067/S0718-07642014000500014

Uso de Minería de Datos en el manejo de Información Geográfica

Arnulfo Castro, Ernesto Sifuentes, Saúl González y Lidia H. Rascón
Universidad Autónoma de Ciudad Juárez, Avenida del Charro 450 Norte, CP 32310, Ciudad Juárez, Chihuahua-México (e-mail: arncastr@uacj.mx, esifuent@uacj.mx, saugonza@uacj.mx, lrascon@uacj.mx)

Recibido Dic. 16, 2013; Aceptado Mar. 17, 2014; Versión final recibida May. 6, 2014

Resumen

En este trabajo se presenta una metodología para el manejo de datos sobre información geográfica. La metodología se aplicó en la identificación de materiales de construcción de calles y avenidas de un área geográfica específica de Ciudad Juárez en México. Los datos necesarios para el desarrollo del modelo fueron obtenidos del Sistema de Información Geográfica de la Universidad Autónoma de Ciudad Juárez, los cuales incluyeron la intensidad y las coordenadas X,Y, Z de unos 878 mil registros de la ciudad. El desarrollo se realizó en varias etapas. La primera corresponde al pre-procesamiento y limpieza de los datos; la segunda es la formulación del modelo e incorporación de los datos a un gestor de base de datos; la tercera es el análisis mismo con técnicas de minería de datos utilizando el software Waika to Environment Knowledge Analysis (WEKA). Para la predicción se utilizaron los algoritmos k-means y Make Density Based Clusters. Los resultados fueron contrastados con imágenes satelitales de las zonas analizadas, validando así la metodología propuesta.

Palabras clave: minería de datos, modelado, agrupamiento, k-means, make density

Use of Data Mining in Managing Geographical Information

Abstract

A methodology for the management of data on geographic information is proposed. The methodology was applied to the identification of construction materials of streets and avenues of a defined geographic area in Ciudad Juarez, Mexico. The necessary data were obtained from the Geographic Information System of the Universidad Autónoma de Ciudad Juárez which included the intensity and the coordinates (X, Y, Z) from about 878 thousand records of the city. The development of the method was done in several steps. The first is the pre-processing and cleaning of data; the second step is the model formulation ant its inclusion in a database manager; the third one is the analysis itself with data mining techniques using the WEKA software. The k-means and Make Density Based Clusterer algorithms were utilized for prediction. The results were compared with satellite images of the areas studied, validating in this way the proposed methodology.

Keywords: data mining, modeling, clusters, k-means, make density

INTRODUCCIÓN

Existen diferentes formas de adquirir datos geográficos que se utilizan en los Sistemas de Información Geográfica(SIG). Una de estas técnicas es el Light Detection and Ranging (LIDAR), este método obtiene datos topográficos que son utilizados en diferentes procesos tales como: comparación de sub-áreas, servicios de localización, administración de zonas costeras, análisis de composición de suelos, ubicación geográfica por Internet (Mennis y Guo 2009). Los datos generados por el proceso LIDAR pueden ser manejados por software especializado en los SIG. Este trabajo presenta una forma de hacer uso de los datos obtenidos por LIDAR y procesarlos mediante herramientas de software libre y obtener información a partir de los datos proporcionados.

Procesar datos no es nuevo y la utilización de herramientas de cómputo permite obtener conocimiento de los datos superando problemas de complejidad, volumen, relaciones propias y externas entre los mismos datos. Una de las áreas de investigación que obtiene información de grandes volúmenes de datos es la Minería de Datos y esto ha permitido el desarrollo de esta área de investigación respecto de la capacidad de predicción en base al procesamiento de cantidades masivas de datos.

Se ha utilizado información geoespacial en varios estudiosregistrando que entre el 5 y 20% de todas las consultas en las páginas Web han requerido información geográfica geoespacial(Himmelstein 2005; Kamvar y Baluja 2006; McCurley 2001; Sanderson y Kohler 2004). En particular, una gran cantidad de información geográfica es utilizada y generada por varias tecnologías (Chung, et al., 2011)en donde se aplican a servicios de localización base (*local-basedservices LBS*) y tecnologías de localización consiente con la finalidad de marcar posiciones geo-referenciadas en imágenes que son utilizadas en la Internet (*location-awaretechnologies LAT*). Por otro lado, la minería de datos se ha utilizado en imágenes geoespaciales con la finalidad de supervisar las posibles amenazas en ecosistemas aplicando algoritmos de clusterización K-meansrealizando un análisis sobre la vegetación en las localizaciones geográficas y buscando posibles cambios en el campo o los bosques(Tran, et al., 2011); así mismo se aplicarón técnicas de minería de datos espacial combinada con sistemas de información geográfica para analizar los cambios en bosques durante dos décadas encontrando un 5.28% de deterioro en el bosque en 20 años (Jayasingheet,al., 2013); De igual forma aplicaron técnicas de minería de datos e información geográfica en el análisis de cambio climático y su impacto (Ganguly y Steinhaeuser, 2008);También se ha aplicadoel algoritmo de clusterización k-means en el análisis de patrones espaciales embebidos en bases de datos espaciales (He y Jiubin, 2010); otra aplicación fue la combinación de los algoritmos de clusterizacion k-means con Expectation-Maximization (EM) y un método simple de proyección para la visualización y reducción de datos (Dogdas y Akyokus 2013);En tanto que el algoritmo de clusterizacion k-means y se ha implementado en un entorno distribuido estableciendo relaciones entre datos no estructurados sobre patrones de comportamiento (Anchalia et. al.,2013); De igual forma se han aplicado técnicas alternativas en el procesamiento de imágenes geoespaciales tales como el sistema de monitoreo a escala continental con resolución moderada y el sistema de alta resolución que hace uso de sobre vuelos del *AerialDetectionSurvey*(ADS); Así como también se ha aplicado la minería de datos en los sistemas de información geográfica marítimos (Li et. al., 2010)

Por otro lado, se ha hecho uso de información geoespacial en sistemas de seguridad, en donde se utiliza el Descubrimiento de Conocimiento Geográfico (*Geographic Knowledge Discovery GKD*) con la finalidad de detectar patrones de comportamiento. En esta técnica se analizan datos de crímenes que son contrastados con datos socio-económicos y socio-demográficos que determinan patrones de co-distribución y contribuyen en la formulación de crímenes(Phillips y Lckjai 2012)de tal forma que es posible crear unanálisis inteligente del crimen(Wortley y Mazerolle, 2008) con los algoritmos implementados en la minería de datos ya que se pueden identificar los patrones de comportamiento criminal,lo cual coadyuva en el cómo, cuándo y donde ocurren ciertos crímenes.

Otra área que utiliza datos geográficos mediante la minería de datos es el análisis de trayectorias (Bongorny et al., 2011) y se ha implementado en un software de trayectoria semántica en WEKA-STPM que opera con bases de datos construidas bajo las especificaciones de la OGC(Bogorny´s Vania 2013) y se aplica para la toma de decisiones de acuerdo a las trayectorias registradas (Bongorny et al., 2009).

DESARROLLO

La figura 1 muestra una representación a bloques de las etapas y procesamiento de datos realizados para obtener información de las estimaciones requeridas en el presente trabajo. Algunas estimaciones son: porcentaje de casas con techo de concreto, porcentaje de avenidas o calles con asfalto.

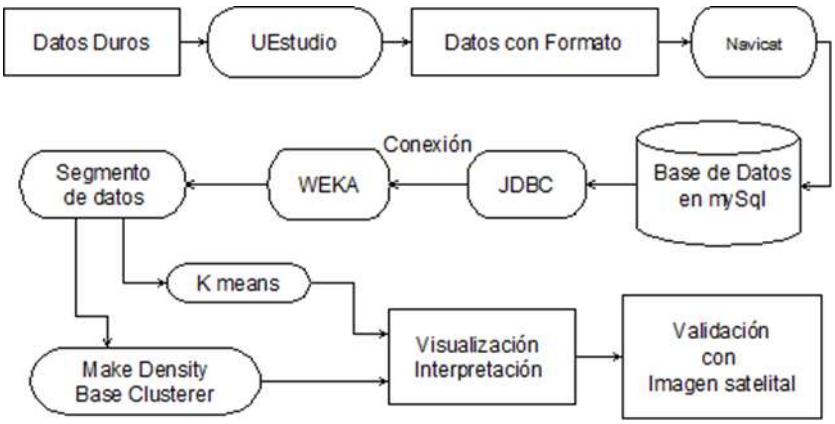

Fig. 1. Diagrama completo del proceso aplicado.

De los elementos presentados en la figura1, los rectángulos representanlos datos con diferentes propiedades y las semi-elipses representan el software que se utilizópara procesar dichos datos. El proceso inicia con los datos duros, denominados así porque no tienen formato alguno, y se obtuvieron del Sistema de Información Geográfica perteneciente a la Universidad Autónoma de Ciudad Juárez (UACJ).Se recibieron 878,512 registrosy se procesaron con el software UEStudio v13.0, debido a que este software permite insertar columnas verticales y separar elementos en archivos de tipo texto. Para nuestro proceso se separaron los registros y se les anexó un identificador. Los datos con formato se presentan en la figura 2, incluyendo el Id que se agregó y que formo parte de la estructura de la tabla “Datosweka” que se muestra

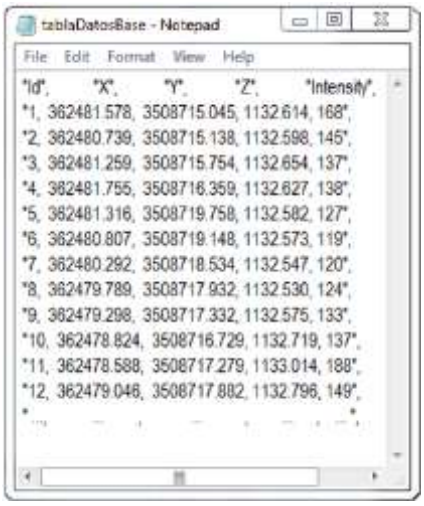

Fig.2. Datos proporcionados por el SIG de la UACJ.

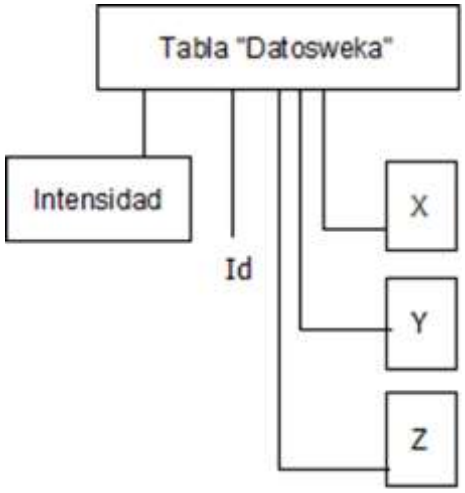

Fig.3. Datos básicos de la tabla Datosweka.

El procesamiento de los datos en un contexto permite obtener cualquier tipo de información deseada. Además, los datos pueden ser casi de cualquier tipo: numéricos,texto, alfanuméricos, ondas electromagnéticas, geo-posicionales, etc. Dentro de las alternativas de software, el más adecuado para manejar datos es el Sistema Gestor de Base de Datos (SGBD) comúnmente conocido como servidor, este software permite administrar los datos y obtener rangos específicos de los mismos con múltiples filtros. Se eligióel MySql 5.0 como SGBD para administrar los datos que se utilizan en este procesamiento.

Se utilizó la interfaz de usuario Navicat 8.0 para crear la base de datos, la tabla que contiene los datos y el proceso de importación de los mismos al servidor de MySql. El servidor permite la comunicación conla herramienta de minería de datos WaikatoEnvironmentKnowledgeAnalysis (WEKA) mediante un intermediario que en este caso fue Java Data Base Conector (JDBC), el cual fue configurado y probado previamente.

Lograda la comunicación, se determinó dividir los 878,512 registros en 16 segmentos lógicos de 54907 registros cada uno.Esta división se realizó para facilitar el procesamiento de los registros por parte del software de minería de datos WEKA, se realizó una prueba de procesamiento con 320,000 registros y el computador no logró generar resultados debido a la saturación de la memoria de almacenamiento.La figura 4 presenta la división de segmentos que fue utilizada.

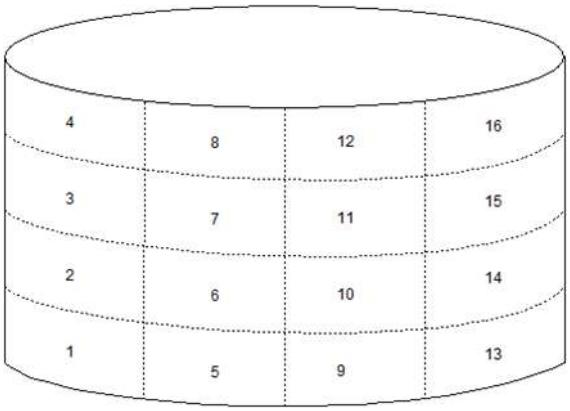

Fig.4. Segmentación de la base de datos.

Una vez creados los segmentos de registros y almacenados en el servidor, se selecciona el segmento de datos mediante consultas en el lenguaje estructurado de consultas(SQL).

WEKA ha implementado diversos algoritmos con distintos propósitos para el área de clústeres de los que destacan:Cobweb, ExpectationMaximization (EM), DensityBasedSpatialClustering of ApplicationswithNoise (DBSCAN), k-means, y MakeDensityBasedClusters. Se seleccionaron los algoritmos K-meansy MakeDensityBasedclusters: el primero debido a su buen desempeño en la creación de clasificaciones de alta similitud entre miembros y baja similitud (Kamber J., 2006) entre grupos, en donde la semejanza de los grupos está determinada por la media de sus miembros;el segundo debido a que realiza funciones similares al anterior pero con la diferencia de que aplica dos algoritmos en su procesamiento para determinar los grupos en zonas altamente pobladas (Ian E. 2005).Así mismo, es posible aplicar los dos algoritmos seleccionados en forma embebida, es decir, los resultados que arroja el primer algoritmo son utilizados por el segundo algoritmo y viceversa.

RESULTADOS

En WEKA se implementaron las consultas de segmentos y se aplicaron los algoritmos de clasificación elegidos, el resultado de aplicar el algoritmo k-means con dos clústeressobre los datos del segmento 2 puede apreciarse en la figura 5, de igual formase aplicó al segmento 9 el algoritmo MakeDensityBasedClusters con cinco clústeres , el resultado de dicho procesamiento se aprecia en la figura 6 en la cual se posible distinguir de manera visual varios elementos como calles, techos y estacionamientos.

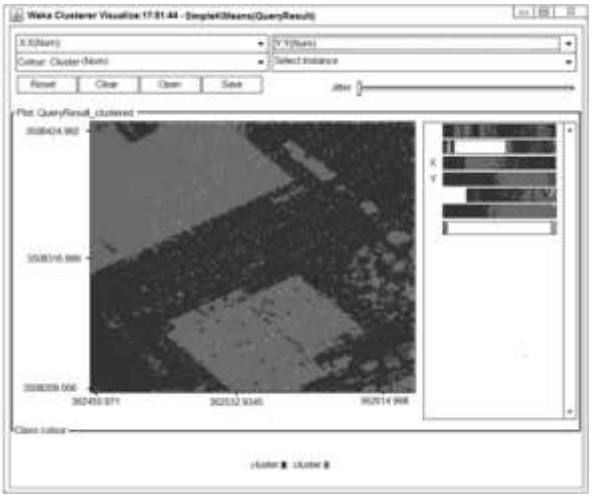

Fig.5. Segmento 2con algoritmo K-means y 2 Clusters.

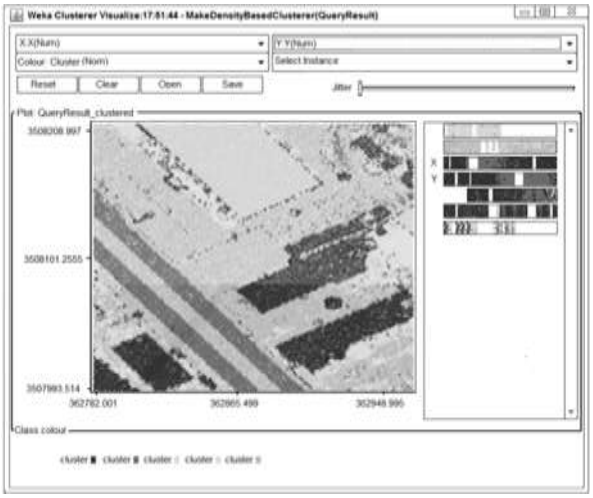

Fig.6. Segmento9 con algoritmoMakeDensityBasedClustersy 5 clusters.

El resultado obtenido figura 5 muestra que con sólo dos elementos de clasificación se agrupan múltiples elementos que representan materiales diferentes que impiden reconocer visualmente elementos como carretera o edificios; por otro lado el resultado de la figura 6 permite reconocer elementos como la carretera, techos, estacionamientos, y calles que en la figura 5 no es posible reconocer.

Con la finalidad de validar estos cálculos se procedió a realizar una comparativa entre las aproximaciones obtenidas mediante WEKA y una foto satelital sobre un área específica ubicada Ciudad Juárez Chihuahua. Los resultados obtenidos del cuadrante 2 después de aplicar el –algoritmo K-means con 2 clústeres se muestra en la figura 7. El resultado se contrastó con la foto satelital figura 8, lo que permite realizar una comparativa visual.

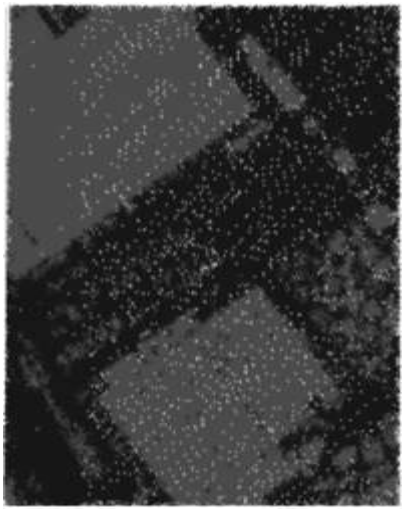

Fig.7: Segmento 2 con algoritmo K-means y 2 clústeres.

Fig.8: Fotografía satelital correspondiente al segmento 2.

Aplicar el algoritmo K-mean al segmento 2 permitió determinar las incidencias de los patrones entre los diversos materiales, el resultado presentado permite determinar que los tonos claros representan techos de lámina y los oscuros áreas asfaltadas en contraste con la fotografía satelital. Delacomparativa de resultados (figuras 7 y 8), se logró determinar que el 53% del segmento representa calles, caminos, estacionamientos, y otros. El complemento de este segmento, 47% representa construcciones como casas y edificios.

Así mismo se le aplicaron al segmento 9 los dos algoritmos con 5 clústeres y estos resultados se compararon con la imagen satelital del mismo segmento. El resultado de la aplicación de los dos algoritmos se puede apreciar en la figura 9 y la imagen satelital correspondiente se presenta en la figura 10.

Fig.9. Representación del cuadrante 9 con técnicas fusionadas y 5 clústeres.

Fig.10. Imagen satelital de referencia cuadrante 9.

Al aplicar los dos algoritmos sobre un mismo segmento con cinco clústeres permite obtener una mejora sustancial ya que presente elementos como calles, estacionamientos o construcciones con una gran semejanza a los objetos reales, lo que se puede apreciar al comparar la figura 9 con la 10. En cuanto a la determinación de porcentajes se registró un 71% para elementos como banquetas, calles y estacionamientos y un 29% para los edificios y casas.

DISCUSIÓN

Se logró adecuar los datos duros con la capacidad de poder accesar a estos desde el software WEKA.

Se logró obtener información de aplicar a datos los algoritmos K-means y MakeDensityBasedClustersy determinar porcentajes de materiales específicos.

Se obtuvo un método que permite realizar selecciones específicas de datos a los cuales se les pueden aplicar algoritmos de agrupamiento que generan resultados gráficos y porcentajes de materiales que permiten realizar predicciones de materiales específicos como asfalto o concreto.

Aplicar los algoritmos de clasificación a los datos geográficos permite obtener información relevante sobre la cantidad de elementos que se sub-clasifican tales como el porcentaje de concreto en una sección de la ciudad, así mismo también se puede realizar una aproximación sobre la cantidad de casas con techo de laminado o el porcentaje de calles pavimentadas por mencionar algunos de los cálculos que se pueden realizar a través del análisis de los datos con ésta herramienta de la Minería de datos.

CONCLUSIONES

Utilizando herramientas de software libre es posible aplicar algoritmos de minería de datos para extraer información relevante de una gran cantidad de información. En este trabajo se presenta una metodología que utiliza los algoritmos K-means y MakeDensityBasedClustersdel software WEKA en el manejo Información Geográfica. Lo cual permitió identificar diversos materiales de construcción, el porcentaje de avenidas y calles con asfalto en un área geográfica especifica. Es decir, se pueden realizar predicciones sobre materiales de construcción únicamente con su ubicación geográfica y un dato de intensidad de material.Como trabajo a futuro se puede crear un mecanismo que automatice el proceso de cálculo de un área específica a múltiples áreas y calcule el porcentaje de materiales de cada edificio o carretera. La metodología propuesta en el presente trabajo puede ser utilizada en otras áreas de aplicación.

REFERENCIAS

Anchalia, P.P.; Koundinya, A.K.; Srinath, N.K., "MapReduce Design of K-Means ClusteringAlgorithm," *Information Science and Applications (ICISA), 2013 International Conference on* , vol., no., pp.1,5, 24-26 (2013).

Bongorny, V., Bongorny´sVania., http:/www.inf.ufsc.br/~vania/software.html, acceso 12 de junio(2013).

Bongorny, V., B. Kuijpers, y L.O. Alvarez.,*St-dmql: A semantic trajectory data mining query language***,** Interntaional Journal of Geographical Information Science 23, 1245-1276, (2009).

Bongorny, V., H. Avancini, B. Cesar de Paula, C. Rocha y L.O. Alvares, *Weka-STPM: A software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization*, Transactions in GIS volume 15 section 2, 227-248, (2011).

Chung L., Y. Hsin, W. Shih, *An image annotation approach using location references to enhance*, Expert Systems with Applications 38, 13792-13802, (2011).

Dogdas, T.; Akyokus, S., "Document clustering using GIS visualizing and EM clustering method," *Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on* , vol., no., pp.1,4, 19-21 (2013).

Ganguly, A.R.; Steinhaeuser, K., "Data Mining for Climate Change and Impacts," *Data Mining Workshops, 2008. ICDMW '08. IEEE International Conference on* , vol., no., pp.385,394, 15-19 (2008).

He Bing Quan; Jiubin Wang; Chao Li, "The Research of the Data Mining Based on the Spatial Database Technology," *Information Science and Management Engineering (ISME), 2010 International Conference of*, vol.2, no., pp.203,206, 7-8 (2010).

Himmelstein, M., *Local search: The Internet is the yellow pages*, IEE computer, 26-34, (2005).
Ian, E., *Data Mining Practical Machine Learning Tools and Techniques*, Waikato: Elsevier, (2005).

Jayasinghe, P.K.S.C.; Yoshida, Masao, "Spatial data mining technique to evaluate forest extent changes using GIS and remote sensing," *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on* , vol., no., pp.222,227, 11-15 (2013).

Kamber, J., *Data Mining Concepts and Techniques*. San Francisco: Elsevier,(2006).

Kamvar, M., S. Baluja, *A large scale study of wireless search behavior: Google mobile search*, In Proceedings of the SIGCHI conference on human factors in computing systems, 701-709, (2006).

Li Gaixiao; Peng Rencan; Zheng Yidong; Zhao Jidong, "Spatial Data Mining and its application in Marine Geographical Information System," *Environmental Science and Information Application Technology (ESIAT), 2010 International Conference on* , vol.1, no., pp.514,516, 17-18 (2010).

McCurley, K.S., *Geospatial mapping and navigation of the web*, In Proceedings of the 10th international conference on world wide web Hong Kong, 221-229, (2001).

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

Onshre LIDAR Data Sets, Journal of Coastal Research, 19-29,(2011).

Tran, R., M. F. Hoffman, K. Jitendra, W.W. Hargrove, Cluster Analysis-Based Approaches forGeospatio-temporal Data Mining of Massive Data Sets for Identification of Forest Threats, International Conference on Computational Science ICCS Procedia Computer Science 4, 1612-1621, (2011).

Phillips, P., L. Lckjai, *Mining co-distribution patterns for large crime datasets*, Expert Systems with Applications 39, 11556-11563, (2012).

Sanderson, M., J. Kohler, *Analyzing geographica queries, In Proceedings of the workshop on geographic information retrieval*, Sheffield UK, 25-29,(2004).

Wortley, R., L. Mazerolle, Environment criminology and crime analysis, Willan Publishing, (2008).