Sistema Geográfico para Vigilancia Epidemiológica

Informe Final

Proyecto de Grado

Departamento de Investigación Operativa – IN.CO.
Facultad de Ingeniería - UDELAR

Octubre 2008

Tutores:
Ing. Omar Viera (Fing)
Ing. Leonardo Loureiro (Ica)

Daniel Cejas 2.839.367-3
Nicolás Cestari 4.150.467-3

Resumen

El siguiente informe presenta de forma detallada la solución obtenida para el presente proyecto de grado. El problema planteado fue la investigación e implementación de un grupo de herramientas para el estudio y el análisis orientado a la vigilancia epidemiológica sobre un Sistema de Información Geográfica. El método utilizado fue iterativo e incremental y las herramientas desarrolladas se basaron gran parte en funcionalidades de programas existentes de epidemiología creados a fines de los 90 y principios de la década del 2000 como son Epiinfo y SIGEpi, y en el asesoramiento de especialistas en el área. El resultado final es un conjunto de instrumentos epidemiológicos para uso especifico de ArcGIS que cumplen con la mayoría de los puntos planteados; Esto es la generación de algoritmos para el tratamiento de datos estadísticos como ser I-Morans y Pearson, específicos de epidemiología como son Cohortes y Caso Control y de Minería de Datos como Árboles de Decisión, *Knn* y *Kmeans*.

Palabras claves: Vigilancia Epidemiológica, Minería de Datos, SIG, ArcGis, Epidemiología, Alerta Epidemiológica.

Índice

Capítulo 1

Introducción

Desde principios del siglo XIX, y con el advenimiento de pandemias que diezmaron más de la mitad de la población de Europa (como fueron la peste o la poliomielitis) se comenzó a tomar muy en serio el estudio de las epidemias, sus causas y métodos para minimizar sus consecuencias o prevenir sus apariciones.

Personajes como Louis Pasteur[1] que en 1885 aplicó con gran éxito la primera vacuna contra la rabia a un ser humano, abriendo el camino en esta área; o como el Dr. John Snow[2] que en 1854 utilizo por primera vez una cartografía como una herramienta muy potente para el estudio del comportamiento del cólera, señalando en un mapa del distrito de SoHo de Londres los casos de la enfermedad y tomas de agua potable, lo cual le permitió localizar con precisión una toma de agua contaminada como fuente del brote del mal. Snow, sin saberlo, fue un precursor en el uso de Sistemas de Información Geográfica (SIG) aplicados a la medicina.

El profesor David Rhind define a los Sistemas de Información Geográfica (en adelante SIG) en los siguientes términos[3]:

"*Es un sistema de hardware, software y procedimientos, diseñados para soportar la captura, el manejo, la manipulación, el análisis, el modelado y el despliegue de datos espacialmente referenciados (georeferenciados), para la solución de los problemas complejos del manejo y planeamiento territorial*".

Actualmente el área de SIG aplicado a la Epidemiología esta en franco desarrollo permitiendo hacer toma de decisiones en gran variedad de temas relacionados con Salud Pública como ser desnutrición, enfermedades ocasionadas por causas ambientales e inclusive en casos de desastres naturales como método de contingencia y de logística para llevar asistencia de la mejor forma posible.

Antecedentes

Como ya se ha mencionado antes, existen dos antecedentes a nuestro trabajo: los programas SIGEpi[4] y Epiinfo[5].

SIGEpi es un software desarrollado por la Organización Panamericana de la Salud (OPS) con el propósito de fortalecer las capacidades de análisis epidemiológico de los profesionales e instituciones de salud a través de un conjunto de herramientas de análisis epidemiológico con un valor agregado, como es el uso de SIG para realizar geoanálisis básico. La última versión es la 1.4 del 5 de Marzo de 2005

Epiinfo es un programa de dominio público diseñado por el Centro para el Control de Enfermedades de Atlanta. Permite generar bases de datos y formularios fácilmente, analizar los datos con herramientas estadísticas básicas y representarlos con gráficos y mapas. Actualmente el software es usado como herramienta básica en algunos hospitales uruguayos (como es el caso del CASMU). La última versión es del año 2002.

Contexto

El planteamiento de la necesidad de una herramienta de epidemiología basada en ArcGIS fue dado por la empresa ICA[6] de la cual surgió el presente proyecto.

ICA[6] es una empresa uruguaya que brinda soluciones informáticas de alto nivel de elaboración y especificidad a organizaciones públicas y privadas. Para ello ha orientado sus acciones a la integración de información heterogénea y desarrollo de sistemas basados en tecnologías de última generación.

El marco del proyecto fue la utilización como punto de partida y fuente de consulta internacional los programas similares en el área comentados anteriormente (SIGEpi y Epiinfo) para generar en base al software ArcGis[7] 9.2, distintas herramientas con similares prestaciones pero modernizadas y además agregar otras artes que estos programas carecían (pe. Minería de Datos).

Como consulta local y a fin de obtener información de campo y datos reales para modelar las funciones en el proceso de análisis de las herramientas para poder entender su funcionamiento, se trato de conseguir la conexión de forma informal con expertos del MSP en el área de Vigilancia Epidemiológica. A medida que el proyecto avanzó, el plan de trabajo original se desvirtuó debido a que pese a realizar pedidos formales de asistencia a los jerarcas del Ministerio no se obtuvo respuesta.

Luego, se buscó la colaboración de la Facultad de Medicina a través del Instituto de Higiene y Epidemiología; Esta iniciativa tampoco tuvo eco debido a diferencias de enfoque entre la directora del UVISAP (Unidad de Vigilancia en Salud Pública) y la estructura del proyecto, dado que ella deseaba que fuera realizado enteramente con software libre.

Finalmente, gracias a la colaboración desinteresada de la doctora Rosario Berterretche (con master en epidemiología) que originalmente fue el nexo entre el MSP, la UDELAR y nosotros, pudimos obtener una guía básica para seguir adelante con el proyecto.

Sin embargo, como consecuencia de lo anteriormente descrito, se produjo un atraso en el proyecto por lo que se tuvo que modificar el cronograma así como parte del alcance del mismo.

Definición del problema

Los sistemas usados actualmente para el estudio de epidemiología y que usan SIG como herramientas de geoanálisis (como son EpiInfo y SIGEpi) carecen de un marco específico para este fin. Esto significa que el potencial en la utilización de dichas herramientas se ve afectado porque se utiliza como un elemento de visualización, soporte y corroboración de resultados.

Considerando que actualmente el *software* ArcGIS de ESRI[8] contiene un conjunto de instrumentos especializados en geoanálisis y que además es posible expandirlo con nuevas *toolbox*[I], se pensó en que usarlo como base para nuestro proyecto era una buena elección tecnológica a pesar de que el uso de éste era un requerimiento no funcional, puesto que al contrario de los *software* precedentes tiene la propiedad de combinarse con otras herramientas no especializadas para encontrar por ejemplo, mediante el agregado de minería de datos, información oculta que pueda dar relaciones no evidentes entre los datos estudiados.

Objetivos del proyecto

El trabajo planteado consiste en dos etapas, la primera es la investigación de las distintas tecnologías y herramientas que van a ser necesarias para llevar a cabo el proyecto. Esto se ve reflejado a través de un conjunto de estados del arte.

La segunda etapa consiste en la generación de una serie de herramientas basadas en SIG y Minería de Datos enmarcada en el uso de ArcGIS como *framework* de desarrollo relacionadas con la Salud Pública y epidemiología para que, de forma práctica, y tomando en cuenta datos existentes, se desplieguen alertas o resultados útiles para la toma de decisiones sobre posibles riesgos de la salud en la población como podrían ser epidemias o enfermedades con causas humano-ambientales, como son: pesticidas, desechos químicos, etc.

[I] Una *toolbox* es una forma lógica de agrupar herramientas usadas en ArcGIS para el procesamiento de los datos usados por el programa. Existen *toolbox* para generar simples estadísticas de los datos hasta de complicados geoprocesamientos.

Estado del arte

A fin de entrar en contacto con el proyecto y cada una de sus facetas, se generó un estado del arte con cinco secciones diferentes que se encuentran en un documento anexo:

Estado del arte SIG que engloba las definiciones y características de esta área para poder entender y profundizar en el software base del proyecto.

Estado del arte Epidemiología que permite adentrarse en la disciplina, mostrando como esta organizada, sus principales características, su alcance y los métodos que aplica cuyos resultados son utilizados para realizar toma de decisiones.

Estado del arte de Salud Pública y Minería de Datos pues algunas de nuestras herramientas utilizan este conjunto de técnicas. Se describen sus conceptos básicos, los distintos tipos que existen y se muestran varios algoritmos de los cuales se implementaron tres.

Casos de estudio de SIG, Epidemiología y Minería de Datos el cual intenta realizar una aproximación entre las tres áreas y mostrar cuanto más potente puede ser el uso en forma conjunta de las mismas. Contiene varios ejemplos donde se describen el problema resuelto y una descripción de la solución adoptada.

Estado del arte del software utilizado como referencia el cual usamos como referencia base para realizar la solución a nuestro proyecto.

Método de solución

La idea planteada fue, tomando como base los programas existentes en el área, generar en ArcGIS[7] - ArcMap[9] algunas herramientas con similares características y agregar otras que estos programas carecían.

En cuanto a la codificación, ArcMap permite generar *scripts* en tres posibles lenguajes, Python[10], JavaScript[11] o Visual Basic[12]. El lenguaje utilizado mayormente fue Python, en su versión 2.4, porque no solo es el utilizado por los propios *scripts* del ArcGIS, sino que además posee una curva de aprendizaje baja, es muy potente, tiene buena *performance* y se puede acoplar al software base directamente. Además se utilizó Java para generar una ventana con resultados (ver Índice de Pearson en la página 40).

Testeos

Se realizaron pruebas unitarias para cada modulo creado dado que las herramientas son independientes entre si. Se generaron varios juegos de datos de entrada para cubrir una gran variedad de escenarios posibles.

Resultados obtenidos

Se han generado y analizado[II] una cantidad considerable de herramientas las cuales se pueden agrupar en cuatro categorías:

- Procedimientos específicos de análisis epidemiológicos:
    - Estudios de Cohortes

      Algoritmo que dado una tabla con la información necesaria para cada punto, genera una capa temática con los resultados obtenidos de aplicar cohortes para los mismos.
    - Estudios Caso - Control

      Análogamente con cohortes, a partir de una tabla genera una capa temática con los resultados de aplicar caso-control.
- Procedimientos generales de análisis y estadística de datos:
    - Estadística Descriptiva

      Para los atributos elegidos de una capa temática se le calculan un conjunto de medidas estadísticas de tendencia central y dispersión.
    - Distribución de frecuencias

      Calcula la distribución de frecuencias de los valores de los atributos seleccionados de una capa.
    - Análisis de correlación

      Calcula la matriz de correlación del conjunto de variables o atributos seleccionados para determinar por ejemplo, cuales variables son importantes para un estudio determinado.
- Procedimientos generales para análisis y estadística epidemiológica:
    - Cálculo, estandarización y suavizamiento de tazas

      Permite el calculo de tasas brutas o especificas de grupos de población u otros datos y estandarizarlas si se desea de acuerdo a algún criterio. Se puede usar un método directo o indirecto de estandarización.
    - Identificación de áreas criticas

[II] Aunque algunas herramientas existían originalmente en el software ArcGIS, las mismas se analizaron y probaron para ver su uso y aplicatibilidad ya que las mismas representaban una componente esencial en estudios epidemiológicos.

Permite buscar áreas o regiones que cumplan algún criterio deseado y se pueden usar para localizar índices de salud particulares.

- Construcción de un índice compuesto en salud

  Permite generar un índice compuesto de salud a partir de un conjunto de indicadores seleccionados. Se genera una capa temática con los resultados obtenidos.

- Procedimientos para análisis espacial:
  - Suavizador espacial

    Toma valores de una capa *raster* y "suaviza" los mismos tomando en cuenta la tendencia espacial que tiene un valor respecto de sus vecinos.

  - Valor promedio ponderado espacial

    A partir de una capa *raster* genera a través de un suavizado espacial local definido una capa con valores promedio.

  - Índices de autocorrelación espacial

    Evalúa si ciertos valores siguen una distribución dada o al azar. Permite detectar si hay algún patrón espacial estadísticamente significativo o alguna concentración espacial (*cluster*) de la variable en estudio.

- Procedimientos para descubrimiento de información:
  - Árboles de decisión

    Dada una tabla con los datos, genera un árbol que categoriza la información para hallar relaciones importantes entre los mismos.

  - K-vecino mas cercano (Knn por sus siglas en inglés)

    Algoritmo que, dado un conjunto de puntos definidos con alguna métrica, busca la distancia mínima desde un conjunto de valores de entrada a alguno de los mismos.

  - Agrupación Kmeans

    Agrupa en conjuntos "similares" elementos de un conjunto de valores de entrada.

Conclusiones

El *software* utilizado (ArcGIS) como *framework* de desarrollo en Sistemas de Información Geográfica demostró ser una herramienta versátil y potente. Y, aunque carezca del refinamiento suficiente a la hora de generar estructuras de programación específicas, las *toolbox* demostraron ser lo suficientemente amplias como para realizar herramientas diversas, poderosas y combinables.

Los SIG aplicados a la epidemiología representan una combinación acertada que apoya el análisis de situación de salud, la investigación operacional y la vigilancia para la prevención y el control de problemas de salud. Así mismo, estos sistemas proveen el apoyo analítico para la planeación, programación y evaluación de actividades e intervenciones del sector salud.

Por ello, pueden considerarse parte de los sistemas de apoyo a decisiones para quienes formulan y siguen políticas en salud. Además, representan una nueva tecnología en el campo de la Salud Pública que puede tener múltiples aplicaciones que fortalecen la capacidad de gestión de los servicios de salud.

En cuanto a los objetivos planteados, se cumplieron satisfactoriamente casi en su totalidad. Se generaron cinco partes del estado del arte (SIG, Epidemiología, Minería de datos, Integración de los temas anteriores y Software precedentes), se realizaron un conjunto de herramientas aplicables a epidemiología basados en ArcGIS y se implementaron tres algoritmos de minería de datos (Árboles de decisión, *Knn* y *Kmeans*). El último punto de los objetivos (Alertas epidemiológicas automáticas) no se pudo realizar debido a contratiempos.

Organización del informe

El presente informe consta de varios capítulos con secciones, donde:

- Capítulo 1, contiene la introducción, planteo y definición del problema y cuales son los objetivos planteados.
- Capitulo 2, brinda una definición detallada del problema a nivel de requerimientos generales del producto y específicos del prototipo a generar.
- Capítulo 3, desarrolla en profundidad el análisis, se detalla el alcance proyectado y el logrado, así como el diseño de la solución.
- Capítulo 4, muestra la implementación generada así como las pruebas realizadas y los resultados obtenidos.
- Capítulo 5, contiene las conclusiones y el trabajo a futuro.
- Anexo, contiene el estado del arte de los distintos temas abordados por el proyecto.

Capítulo 2

Descripción detallada del problema

Lo planteado por la empresa ICA[6] a través de nuestro tutor es la generación de un sistema de herramientas que permita el estudio y análisis de datos para uso de profesionales en epidemiología y su posible acceso por gestores en la toma de decisiones que puedan no tener conocimientos particulares sobre la materia.

La idea central es realizar diversas herramientas que abarquen los temas de estadísticas, análisis en zonas geográficas y epidemiológicas integradas en un mismo paquete para aumentar al máximo la comodidad y la potencia de los estudios. A esto se le suma la necesidad de obtener relaciones entre datos almacenados que en un principio están desconectados, puesto que es muy importante para encontrar potenciales causas de ciertas enfermedades.

Finalmente, se busca la conjunción de los puntos anteriores en la generación automática de alertas, iniciadas por valores anormalmente altos de casos de ciertas enfermedades en una zona (epidemia) o de productos potencialmente nocivos para la salud, producidos tanto de forma natural (por ejemplo gases emitidos a la atmósfera por erupciones volcánicas) como por la mano del hombre (por ejemplo emisiones de anhídrido carbónico al aire por los vehículos y fábricas).

Resumen del Estado del Arte

Sistemas de Información Geográfica

Los Sistema de Información Geográfica (SIG) son una disciplina que ha evolucionado y sigue evolucionando con tal rapidez, que una definición de lo que es o de lo que hace cambia a cada momento, hasta el punto que lo único cierto es que cualquier definición que demos de ella ahora ya no será válida dentro de unos años.

En realidad, la definición misma no es tan importante como las ideas básicas que están detrás de un SIG, a saber[13]:

- Ser “geográfico” significa que contiene datos y conceptos relacionados con las distribuciones espaciales.
- Que “información” implica alguna forma de transmisión de datos, ideas o análisis, en general como ayuda para la toma de decisiones.

- Que por ser un "sistema" conlleva una secuencia de entradas, procedimientos y salidas.

Además, los tres elementos arriba mencionados adquieren su funcionalidad basados en un escenario tecnológico actual con las posibilidades que ofrece la alta tecnología.

Por lo dicho anteriormente, el propósito general de un SIG es el de proveer un área de trabajo que permita capturar, almacenar, analizar y manejar información y atributos asociados que están espacialmente referenciados. Esto beneficia, por ejemplo, a la toma de decisiones para el uso inteligente de los recursos y para poder organizar mejor las áreas de uso.

Un SIG puede mostrar información de forma interactiva cuando se tiene en un PC, esto permite brindar datos que no aparecen en mapas impresos: como ser la generación de un río y realizar simulaciones sobre el mismo para ver como influyen las lluvias. La forma que uno elige para analizar y desplegar los datos va a depender de cómo se modelen los objetos de la realidad particular.

Las partes de un SIG

Un SIG es una combinación de personal calificado, datos espaciales y descriptivos, métodos analíticos, y software y hardware de computación organizados para automatizar, manejar y brindar información por medio de una interfase geográfica[14].

Personal Calificado

Cuando se diseña un modelo de datos, se crea un software de aplicación o se escribe documentación de usuario es importante conocer el tipo de usuario al que va dirigido.

Datos espaciales y descriptivos

Un SIG procesa cualquier dato que tenga algún componente espacial. Esta información puede provenir de muchas fuentes diversas, como ser: fotografías aéreas, imágenes satelitales, mapas digitales o datos de colonización.

Procedimientos y Análisis

Los especialistas que manejan los SIG emplean funciones, procedimientos y su propia experiencia personal para generar un análisis sobre los datos.

Software SIG

La idea clave tras el software de SIG es que, de hecho es un sistema manejador de bases de datos geográficas. Las bases de datos geográficas están implementadas directamente sobre sistemas de bases de datos relacionales u orientados a objetos.

La razón de esto es tener todas las propiedades de las bases de datos comunes, como ser backup de datos, definición de tablas, manejo de transacciones y herramientas de administración del sistema; y a esto agregarle las extensiones SIG para obtener capacidades de almacenamiento eficiente de los datos geográficos, producir mapas y realizar tareas de análisis espacial.

Hardware de Computación

Las computadoras vienen en todos los tamaños, desde *Palms* hasta *MainFrames*. Se puede utilizar un software de SIG para casi cualquier computadora.

Epidemiología

Su significado deriva del griego *Epi* (sobre), *Demos* (pueblo), *Logos* (ciencia). Esto podría entenderse como el estudio realizado sobre las poblaciones.

Definición

Una definición técnica es la que propone que la epidemiología es "e*l estudio de la distribución y determinantes de la frecuencia de la enfermedad en poblaciones humanas."* [15]

Las primeras definiciones encontradas se corresponden a la conceptualización surgida en los albores de la epidemiología, cuando ésta centró su interés en el estudio de procesos infecciosos transmisibles (pestes) que afectaban a grandes grupos humanos. Estas enfermedades, llamadas genéricamente epidemias, resultaban en un gran número de muertes, frente a las cuales la medicina de aquella época no tenía nada efectivo que ofrecer.

La evolución científica y tecnológica, y el cambio en el nivel de vida de las poblaciones modificaron el tipo de enfermedades que afectaron en mayor número y más gravemente a las poblaciones. Esta evolución incluyó en el estudio epidemiológico a enfermedades no infecciosas cuya elevada frecuencia de aparición y dispersión no eran consecuencia de los mecanismos clásicos de transmisión conocidos de las enfermedades infecciosas transmisibles. Estas enfermedades son conocidas actualmente

como enfermedades crónicas no transmisibles (NCD por su sigla en inglés) y también son materia importante en el estudio de la epidemiología moderna.

Salud Pública y Epidemiología: diferencias

Es importante destacar la diferencia entre Salud Pública y Epidemiología ya que ambas comparten el interés por el "colectivo" (demos) contando con un conjunto de conocimientos específicos, así como una metodología y aproximación racional que les es característica.

Sin embargo, la Salud Pública se apoya en la epidemiología para enfrentar la salud y sus problemas en una perspectiva colectiva, pero va más allá al preocuparse también de los elementos que conducen a la corrección de situaciones indeseadas mediante la organización, administración y aplicación de medidas efectivas de prevención y control.

Minería de Datos

El Descubrimiento de Conocimiento en Bases de Datos es el proceso de identificar información válida, nueva, potencialmente útil y entendible en los datos. La Minería de Datos va un paso más en el proceso de Descubrimiento de Conocimientos y consiste básicamente en algoritmos que buscan patrones o modelos en los datos.

Las técnicas de Minería de Datos no son nuevas y fueron perfeccionándose gracias a un largo proceso de investigación y desarrollo. Esta evolución comenzó en la década del 70 cuando los datos de negocios fueron almacenados por primera vez en una computadora, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real.

Realizar Minería de Datos requiere más que simplemente aplicar técnicas sofisticadas a un conjunto de datos. Las técnicas usadas en ésta representan a un conjunto de técnicas estadísticas, de reconocimiento de patrones y aprendizaje automático. La Minería de Datos es una actividad de extracción de información cuyo objetivo es descubrir hechos ocultos contenidos en las bases de datos mediante una combinación de aprendizaje automático, análisis estadístico, técnicas de modelado y tecnología de base de datos, a su vez busca relaciones sutiles en los datos y deduce reglas que permiten la predicción de futuros resultados[16].

Definición

Minería de Datos se puede definir clásicamente como la extracción de información oculta y predecible desde grandes bases de datos. Es un

término aplicado al conjunto de técnicas que pueden ser usadas para encontrar estructuras y relaciones subyacentes en volúmenes muy grandes de información o datos.[16]

La etimología del termino Minería de Datos proviene de la similitud con la búsqueda en una mina de minerales valiosos entre material sin valor. En ambos casos se requiere examinar una inmensa cantidad de material, o investigar de forma inteligente hasta encontrar exactamente donde se encuentra la "veta" valiosa.

Requerimientos

El relevamiento efectuado se basa en información obtenida a partir de reuniones con los tutores del proyecto, *papers* y páginas relacionadas, generados por expertos en el área, además de la colaboración puntual de una persona idónea en epidemiología quien accedió a evacuar ciertas dudas importantes cuando la colaboración del MSP y de la Facultad de Medicina fue malograda. A partir de esto se llegó a generar un conjunto de requerimientos funcionales y no funcionales que el producto debería tener. Además se logró sintetizar el esquema básico de proceder de los estudios epidemiológicos que nosotros debimos plasmar en nuestro proyecto.

Funcionales

- Disponer de un conjunto de herramientas aplicables en epidemiología tomando como base los programas ya existentes en la materia y con el asesoramiento de personal especializado en la materia.
- Permitir la generación de alertas automáticas a partir de cierta información obtenida bien de datos crudos o bien como resultado de la aplicación de alguna herramienta.
- Generar algoritmos de minería de datos que interactúen con estructuras y datos epidemiológicos a fin de permitir la clasificación y el descubrimiento de relaciones ocultas en la información existente.

No Funcionales

- Este sistema de herramientas debe ser hecho usando como software base el ArcGIS 9.2 de ESRI[8].
- Obtener una *performance* mínima para las herramientas generadas.

Capítulo 3

Este capítulo plasma todo lo relacionado al análisis y diseño realizado para el sistema geográfico de Vigilancia Epidemiológica.

Análisis

El proceder epidemiológico esta representado en la Figura 1. De la realidad se recogen muestras discretas o discretizadas[III] que completan la base de datos del sistema.

Estos datos alimentan, junto a un SIG[IV] base, a las funciones y procedimientos utilizados por los epidemiólogos. Del uso de los mismos se consigue uno o varios resultados que luego, al interpretarse, permite tomar decisiones en el área de la Salud Pública.

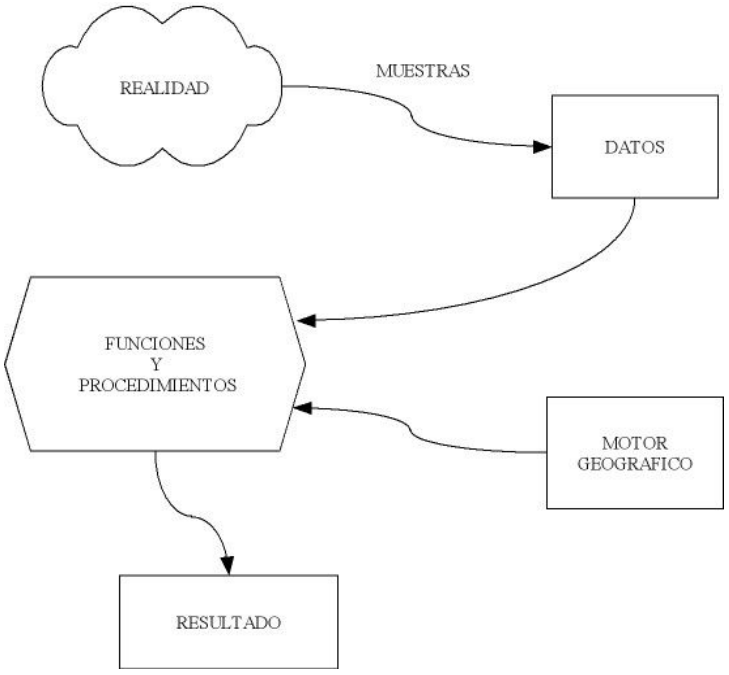


Figura 1 - Esquema básico de funcionamiento de un estudio epidemiológico

[III] Decimos que se discretizan la muestras porque el epidemiólogo para poder hacer los estudios toma valores en puntos específicos del espacio continuo de investigación. Esto sucede porque los modelos utilizados por los mismos tienen entradas discretas.

[IV] En realidad, los estudios realizados por lo general no utilizan la componente geográfica y es en este sentido que el proceder epidemiológico se vería beneficiado al agregar la misma como parte de las entradas del sistema a modelar.

Esta etapa del análisis se centra en la generación de un modelo que permita representar de forma correcta los datos, y la estructura general necesaria para realizar los distintos estudios epidemiológicos. Como punto importante, se pudo llegar a una primera conclusión analítica sobre el modelo tipo de estudios epidemiológicos.

La conclusión es que luego de profundizar en las distintas investigaciones encontradas que hablan sobre el tema epidemiológico, se vio que en los diversos estudios realizados en las mismas no existe un único modelo utilizado. Es más, existe básicamente un modelo para cada estudio realizado. Pensamos que esto se debe a que cada epidemiólogo (más allá de utilizar el mismo "Método Científico") tiene una visión distinta de cómo proceder (que funciones utilizar y como modelar la realidad) ante una investigación.

Además, es evidente que no es posible (por lo costoso) representar en un modelo toda la realidad usada por la epidemiología ya que la misma es potencialmente **todo** el universo.

Esto generó que no se pudiera modelar el "Estudio Epidemiológico" tipo.

Lo que se buscó entonces, fue encontrar el modelo básico que pudiese, si bien no automatizar todos los modelos de estudio (porque descubrimos era impracticable) por lo menos ser capaces de mantener en una misma estructura un diseño que pudiese cumplir con dos objetivos básicos que, pensamos, son esenciales a la hora de realizar un estudio científico-epidemiológico y poder hacer comparaciones del mismo con otros:

a) soportar el almacenamiento de los datos científicos de un estudio epidemiológico cualquiera sea su enfoque y sus metadatos asociados.

b) Poder comparar este estudio con otros realizados anteriormente (datos procesados) y utilizar los datos de un estudio (datos crudos) como parte de las entradas de otro que, en principio, parezca no estar asociado al mismo tema[V].

Basados en estos dos preceptos generamos el modelo esquematizado en la Figura 2. El objetivo del mismo no es el de automatizar estudios epidemiológicos (esto lo harían distintos módulos que trabajarían sobre el modelo) sino el de ser el punto base de donde poder ingresar y obtener datos

[V] Esto tiene como potencial el uso de Minería de Datos sobre datos provenientes de estudios distintos y en principio incompatibles.

de estudios epidemiológicos para poder hacer por ejemplo, minería de datos sobre esa colección y encontrar relaciones entre distinto proyectos realizados que sean, en principio, independientes y almacenados en forma de datos, con lo cual generamos no solo un modelo que cumple los dos objetivos señalados anteriormente sino que como beneficio adicional encontramos el punto que tienen en común los distintos estudios epidemiológicos: Datos y sus metadatos.

Cuando se realiza un estudio epidemiológico (vea Figura 1) lo que uno hace es recabar datos discretizados que luego (siguiendo un Proceder Científico) ingresa a una (o varias) Funciones y Procedimientos que a su vez devuelven una serie de valores que son interpretados directamente por el epidemiólogo y éste por último toma decisiones a partir de sus resultados.

El objetivo "grosso modo" de un estudio epidemiólogo es encontrar las posibles relaciones (de tipo causa-efecto) entre los distintos objetos en estudio. Para esto usa un procedimiento matemático (que puede ser distinto según el tipo de estudio realizado) que genera ciertos índices y los mismos son usados para la toma de decisiones.

Por tanto, el modelo final debe contemplar estas situaciones, que por simples que parezcan resultan en una complejidad absoluta que resultan en la base fundamental a la hora de diseñar un Sistema de Alerta Epidemiológica.

Modelo

Partiendo de la base del razonamiento del punto anterior se puede realizar un primer modelo general que soporte toda la gama de estudios epidemiológicos posible sin la necesidad de generar estructuras que sesguen a un tipo de estudios particular.

Este modelo representa la realidad de la base de datos que se va a implementar para soportar los requerimientos del proyecto.

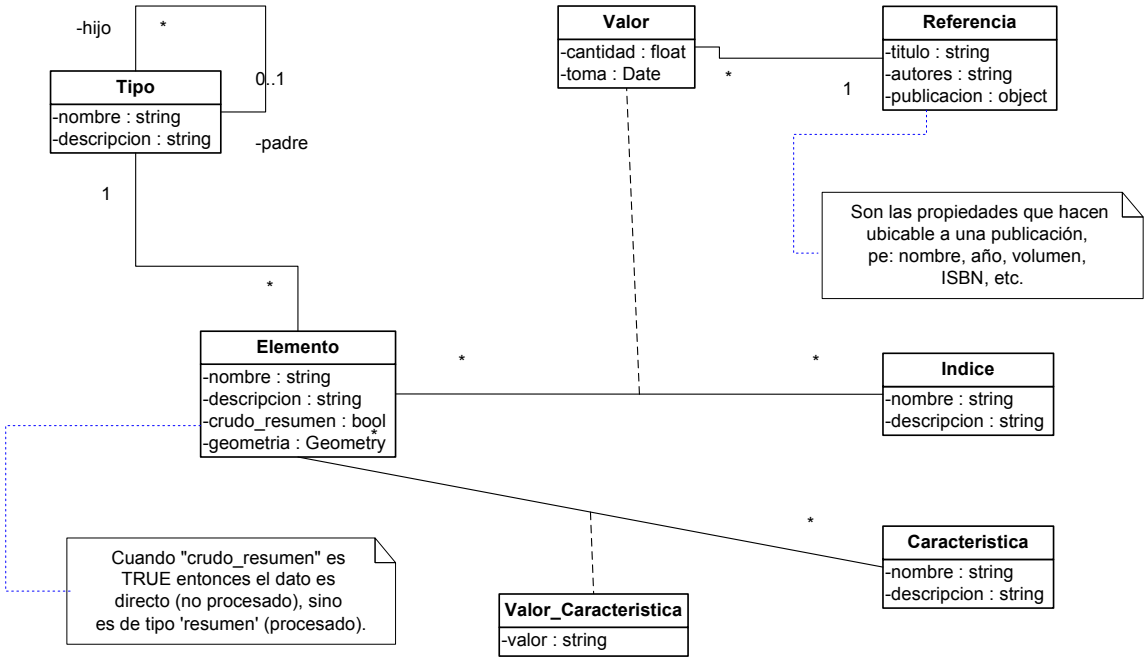


Figura 2 - Modelo Conceptual Básico conteniendo lo que se debería almacenar junto con sus metadatos.

Elemento representa la información de cada objeto en estudio que puede presentarse en forma de un dato crudo o ya resumido (procesado). La "geometría" contiene los atributos geográficos de los datos formado por una geometría que puede representar puntos, líneas o polígonos.

Tipo existe para poder agrupar la información en conjuntos que representen el mismo estudio. Forma una estructura jerárquica para poder recuperar los datos rápidamente.

Índice representa al índice usado para tomar la muestra en el elemento (pe: índice de neutrófilos en sangre, de trílalo métanos en el agua, etc.). Además para cada par Elemento-Índice existe un Valor que es el dato numérico de ese índice para ese elemento.

Referencia es la información asociada al valor y se utiliza como *metadata* para obtener donde y cuando se realizó este estudio particular. La principal aplicación es la de recuperar el origen de los datos usados, y verificar por ejemplo la fiabilidad de los mismos.

Valor contiene el valor asociado a cada índice para un elemento determinado

Característica son las características asociadas al elemento pero que no están directamente relacionadas con los Índices de los elementos. Por ejemplo; el nombre, teléfono, etc.

Como resultado seundario, el modelo permite la incorporación de módulos específicos para estudios epidemiológicos más concretos que necesiten mayor grado de automatización. Además, puede ser usado como base de conocimientos para modelos de simulación.

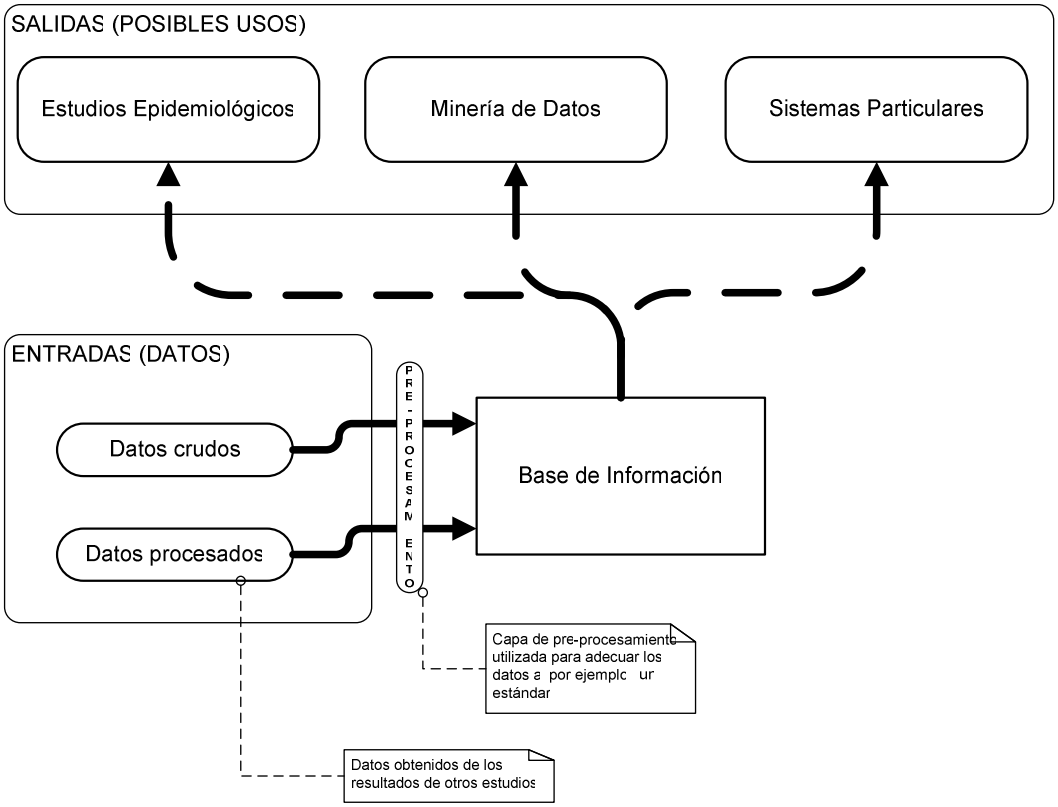


Figura 3 - Modelo de análisis de componentes propuesto

Como vemos en la Figura 3, el modelo propuesto es el representado por el objeto "Base de Información". Cuando un epidemiólogo requiere hacer un estudio (generando datos nuevos o usando resultados de otros estudios), ingresa los datos a la Base de Información a través del módulo de Entrada, en éste podría realizarse un pre-procesamiento para formatear los datos utilizando algún estándar por ejemplo. Como el modelo visto en la Figura 2 es tan general, el epidemiólogo deberá ingresar los datos a conciencia de que van a tener que ser recuperados en algún momento para realizar el estudio en cuestión u algún otro nuevo estudio, por lo cual tiene (a través de metadatos) que agregarle algún tipo de identificación a los mismos. La Base

de Información registra todos los datos usados como entradas para distintos análisis o los resultados de los mismos.

Luego, hay varias opciones para el uso (SALIDAS, vea Figura 3) de los datos de la Base de Información[VI]:

- Pueden ser usados para realizar algún estudio epidemiológico tipo (pe: Caso-Control, Cohortes, etc.).
- Pueden ser ingresados a un motor de minería de datos para descubrir relaciones intrínsecas no visibles[VII] de forma directa.
- Pueden ser usados por otros sistemas para generar nuevos datos o automatizar el uso de los mismos.

En el último punto vemos como por ejemplo los datos de la Base de Información son usados como entrada para sistemas particulares (esto es, por ejemplo: Sistemas de Alertas Tempranas de Vigilancia Epidemiológica) que actúen obteniendo la información de la misma. Además, al tener una componente geográfica, pueden ser usados en simulación (como por ejemplo simular el derrame de un líquido potencialmente peligroso dentro de un río).

Otro uso posible de "Sistemas Particulares" es el diseño de nuevos módulos que funcionen sobre la Base de Información con el objetivo de darle nuevas funcionalidades al sistema.

Como vemos, el potencial es muy grande. El principal problema entonces es diseñar un modelo claro y sencillo para poder ingresar y obtener datos de la Base de Información si queremos que la misma funcione como un modelo de caja negra para que los especialistas en salud (que no tienen por que tener los conocimientos informáticos apropiados) puedan usar el sistema cómodamente.

[VI] Los datos almacenados en la Base de Información pertenecen a varias fuentes de información, por lo que un estudio con datos particulares puede verse beneficiado con datos resultantes de otros estudios para poder comparar o aumentar la precisión del anterior.

[VII] Los datos dentro de la Base de Información están guardados de tal manera que puedan ser rápidamente ingresados como datos de entrada en un estudio de minería de datos.

Protocolos

Existen dos conceptos claramente esenciales para que un sistema con esta característica de "genérico" funcione con el propósito ya especificado. Las *entradas* y las *salidas* y sus formatos juegas uno de los roles más importantes en el tratamiento de la información. En ésta sub-sección se hablará un poco de estas.

ENTRADAS

El protocolo de ENTRADA es bastante directo y depende de la Base de Información. Los datos podrían ingresarse semi-automaticamente de un archivo de tablas (como puede ser una planilla electrónica, o una base .dbf), de una pantalla orientada a usuario, o a través de otro sistema (pe: con un *WebService*[17]). El protocolo de ENTRADA se vería fuertemente beneficiado con un módulo de pre-procesamiento (que separe la Base de Información de las entradas) cuyo propósito principal sea la depuración y estandarización de los datos ingresados.

SALIDAS

El protocolo de SALIDA es un poco más complejo porque se deben dar dos tipos de formato de obtención de datos.

- *Human-Readable*. Este formato permitiría a una persona obtener uno o varios datos para poder realizar procedimientos y formulas[VIII]. Consiste en un metalenguaje de fácil manejo que permita obtener resultados con los datos.

- Servicio Automático. Este formato permite a un sistema externo, conectarse para realizar consultas y así alimentarse de la información cruda o procesada que la base posea. Esto permite además la interconexión entre Bases de Conocimiento distribuidas que se especialicen en distintos temas referentes a la Salud.

[VIII] La idea es que pueda darse un metalenguaje de construcción de formulas que use como datos los de la Base de Información.

Alcance

Alcance proyectado

En una primera instancia, el proyecto estaba enmarcado en la creación de un grupo de herramientas para uso de alerta epidemiológica siguiendo pautas (del MSP por ejemplo) para su construcción. Estas herramientas requerían un software de automatización de procesos de análisis epidemiológico para alertas tempranas.

Además, la empresa ICA estimó conveniente la realización de un prototipo de este grupo de herramientas para verificar la factibilidad del uso del software ArcGIS mediante la creación de una serie de *toolbox* específicas para su uso en epidemiología.

Asimismo, para generar una mejora con respecto a otros *softwares* preexistentes en el área, se previó la generación de *toolbox* en minería de datos.

Alcance logrado

A medida que el proyecto avanzó, como se puede ver en el *Gantt*[18] de la Figura 4, los tiempos se fueron dilatando por diversos motivos lo cual generó la necesidad de recortar el alcance en algunos aspectos que se consideraron menos relevantes.

Se modificaron los requerimientos de que herramientas epidemiológicas realizar debido a que no fue posible contar con el apoyo del MSP o algún cliente que brindara requerimientos y sus respectivas validaciones para las mismas. En cambio se decidió, siguiendo nuestro propio criterio, seguir la línea de los programas preexistentes para obtener dichos requerimientos además de consultar material apropiado cuando se desconocía el uso de alguna herramienta de dichos programas.

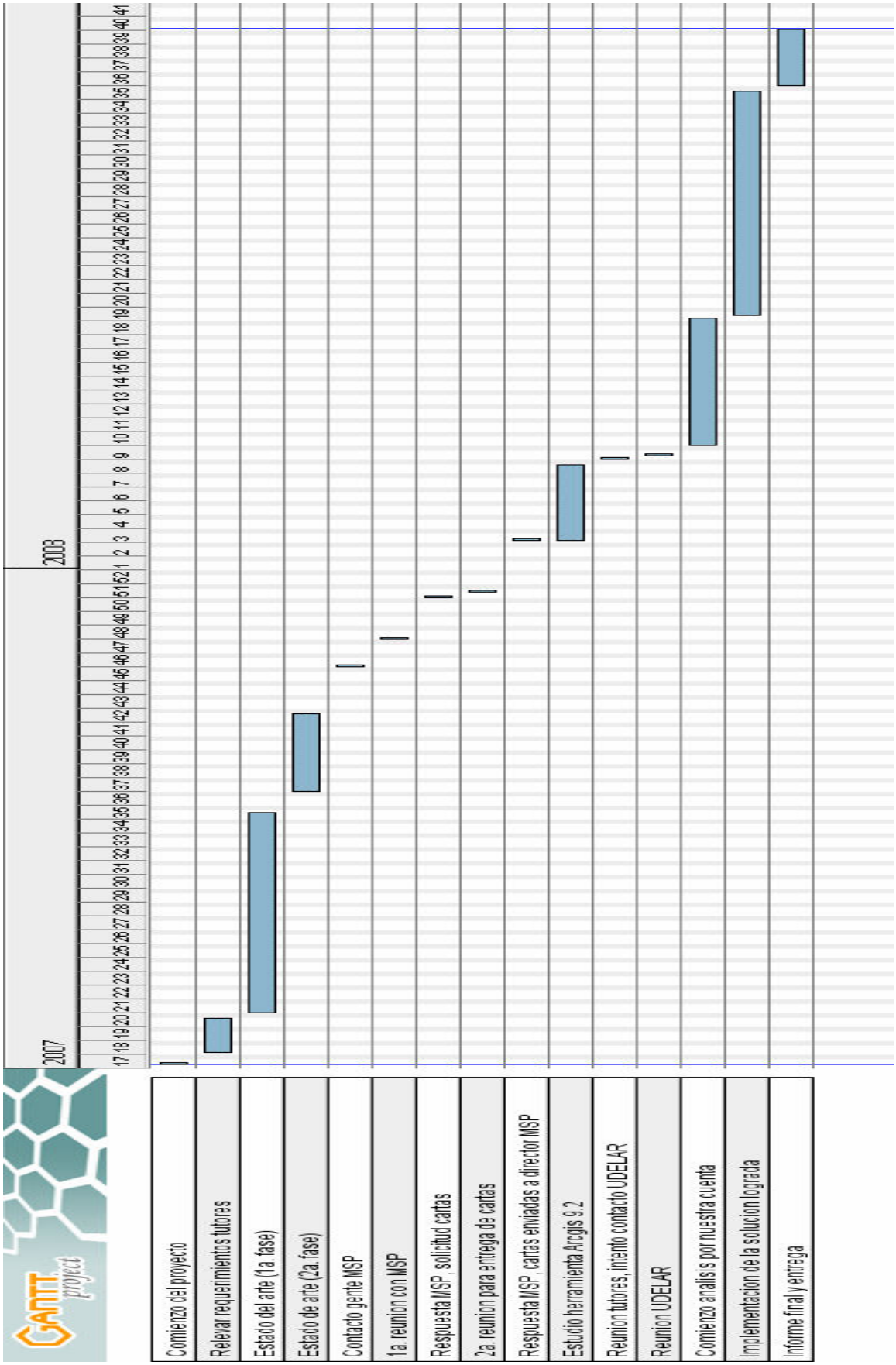


Figura 4 - Gantt conteniendo los tiempos llevados en cada tarea del Proyecto

Por otro lado, nuestro sistema contempla totalmente dos de los tres puntos pedidos, el primero es la generación de herramientas básicas utilizando como ambiente básico de desarrollo el ArcGIS 9.2 y el segundo es la generación de instrumentos basados en minería de datos. El recorte real se realizó al generar un Sistema de Alerta Epidemiológico que no es automático. Sin embargo se pueden (y se hicieron) simulaciones "a mano" para generar un conjunto de datos que podría ser usado como entrada para un sistema de Inteligencia de Negocios por ejemplo, para generar alertas tempranas a gerentes en salud.

Diseño

En el análisis de los requerimientos se estableció que existía un requerimiento de uso de la herramienta ArcGIS. Como se vio hasta ahora, en el análisis del modelo epidemiológico se conocieron los detalles directamente relacionados con epidemiología pero no aquellos requerimientos no funcionales, la razón de esta decisión es que se quería comprender totalmente el tema a desarrollar antes de realizar requerimientos que no estuvieran directamente relacionados con el mismo. Por esta razón se pensaron en dos posibles soluciones de diseño.

Por un lado, se podría implementar un conjunto de herramientas específicas para epidemiología y luego adaptar el motor ArcGIS para que funcione con estas. Y, por otro, se podría usar el ArcGIS como motor básico y construir distintas herramientas sobre el mismo, y así aprovechar todo el potencial y las características que pudiese proveer ArcGIS.

Diseño de Prototipos

Se probaron distintos prototipos y se encontró que era muy engorroso (sino imposible) acoplar ArcGIS a otras herramientas para darles a éstas últimas la potencialidad de análisis requerido. Además, realizar este trabajo significaba simplemente agregarle funcionalidad a aplicaciones existentes (como son SIGEpi y EpiInfo) usadas como modelo de desarrollo y éste finalmente no era un objetivo del proyecto.

Por otro lado, usar el ArcGIS como plataforma de desarrollo permitía utilizar directamente funciones de análisis geoespacial, así como utilizar las características de presentación y despliegue de las soluciones del propio ArcGIS, propiedades deseables para un software de este tipo. Además, el ArcGIS posee un conjunto de librerías para distintos lenguajes conocidos, como son: Python, JavaScript y VisualBasic lo que permite incluirlas fácilmente en distintas soluciones implementadas para el proyecto.

Como resultado, se decidió usar la segunda opción y se implementaron funcionalidades de epidemiología y minería de datos sobre este *software*.

Diseño Realizado

El uso de ArcGIS facilitó la realización de muchos puntos críticos vistos en el análisis ya que esta herramienta tiene solucionados los problemas de entrada y procesamiento de datos, así como el formateado y presentación con reportes de la salida de la solución.

La Figura 5 muestra el diseño del modelo de análisis usando ArcGIS como motor básico de desarrollo. Los datos de entrada serán ingresados con la aplicación ArcCatalog[19] incluido en el *software* ArcGIS.

Dado el diseño inicial de la Figura 3 se ven que las modificaciones realizadas básicamente fueron sobre las entradas que ahora son manejadas por el ArcCatalog, la "Base de Información" fue complementada con distintos componentes ArcGIS al que se le agregaron herramientas especificas de epidemiología y Minería de Datos. Aprovechando la potencialidad que tiene el *software* base, se ve que las salidas pueden ser formateadas por el mismo. Los datos de salida podrían directamente ser cargados a otro sistema más sofisticado de alerta epidemiológica.

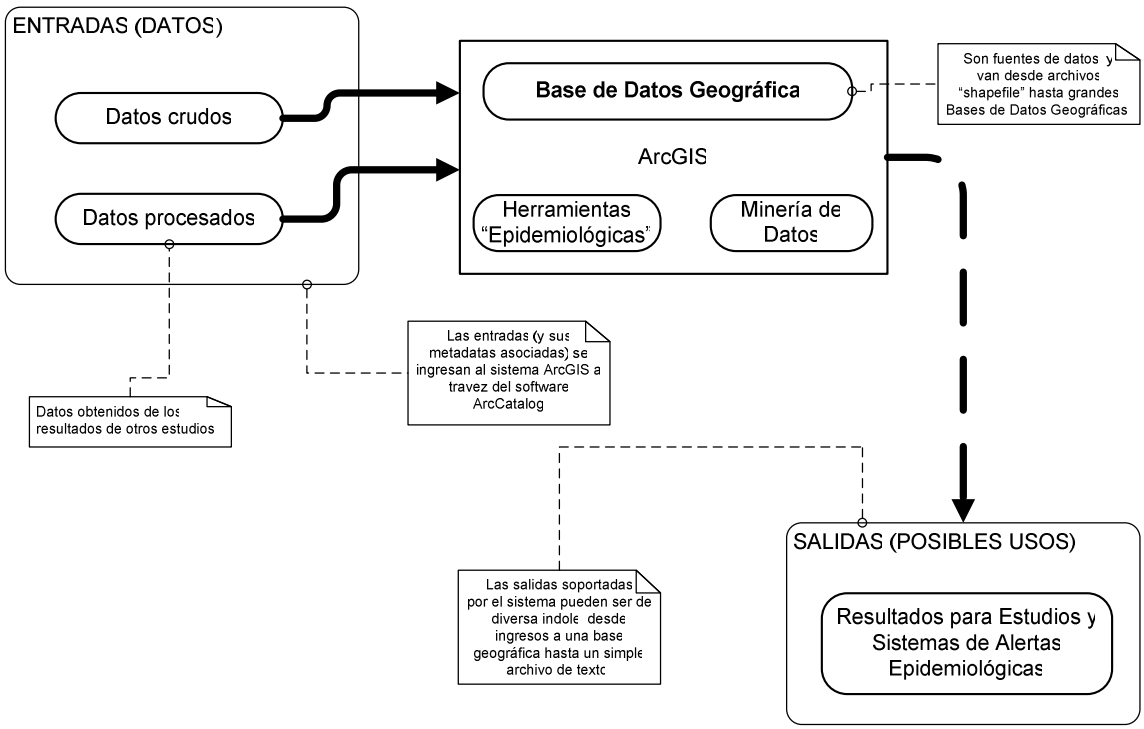


Figura 5 - Modelo de diseño básico utilizando el ArcGIS como Framework de desarrollo

La aplicación de ArcCatalog (Figura 6) organiza y maneja toda la información del SIG como mapas y datos de diversas fuentes, así como modelos de datos y sus metadatos.

Incluye, entre otras, herramientas para:

- Búsqueda de información geográfica desde diversas fuentes de datos.
- Creación, visualización y manejo de metadatos.
- Definir, exportar e importar esquemas de *geodatabases*.

Un usuario utiliza éste software para organizar y encontrar datos geográficos así como no geográficos, o usar y modificar metadatos basados en normas internacionales o preestablecidas localmente. Además, se puede usar ArcCatalog para definir y construir *geodatabases*.

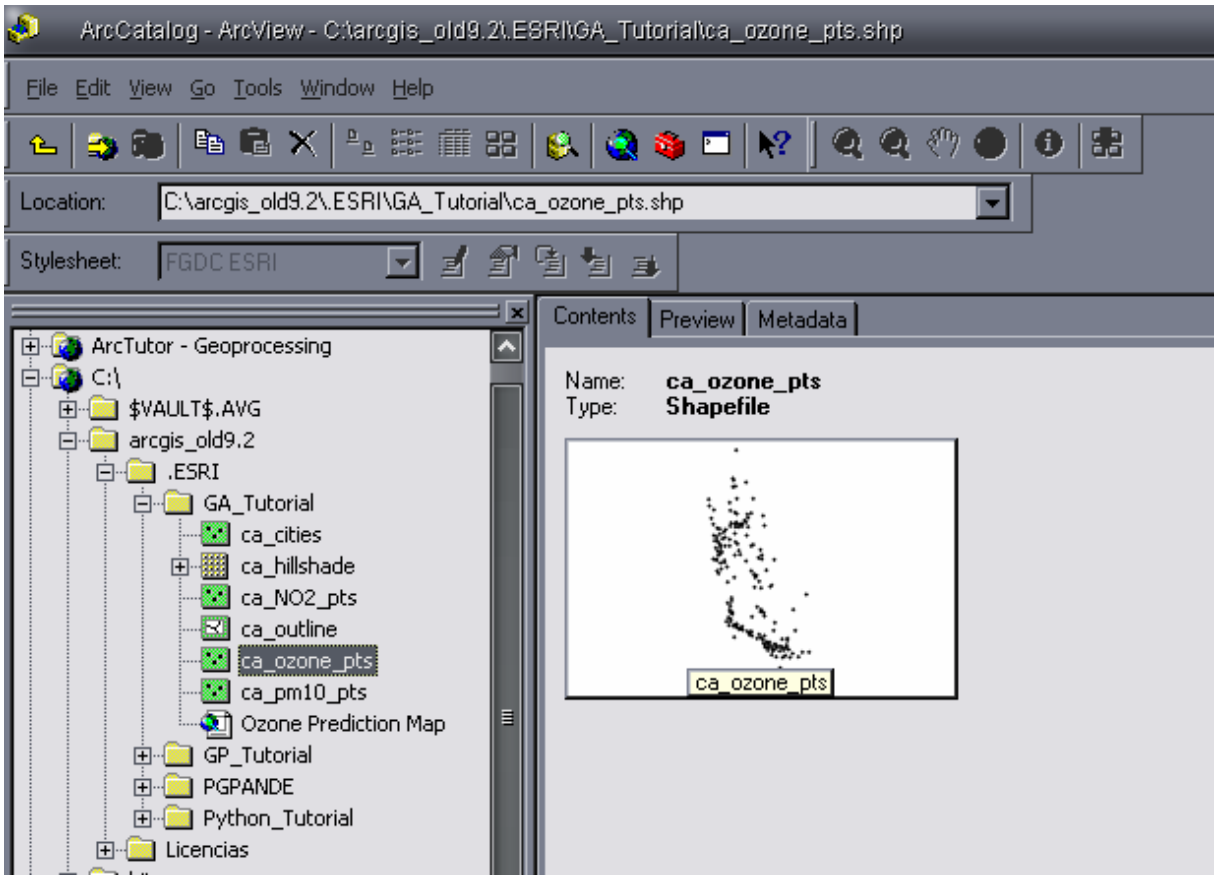


Figura 6 - Visualización del ArcCatalog para manejo de datos de diversa índole (incluso formatos de SIG estándar)

En la Figura 7 vemos la forma en que ArcCatalog trabaja con los metadatos, en particular su pre-visualización. La misma (así como su modificación) puede seguir un estándar establecido.

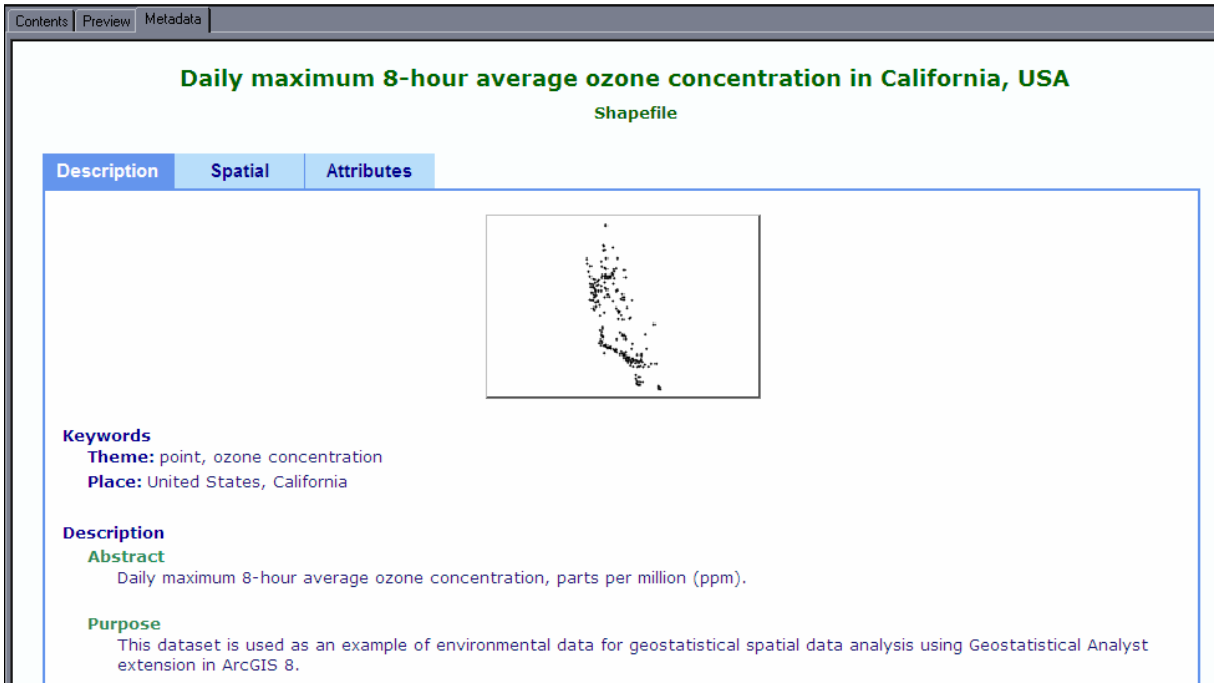


Figura 7 - Se ve la descripción del elemento (esta metadata puede expandirse al tipo de metadato estándar que se desee, pe: FGCD metadata estándar, ISO-19139, etc.)

Además, el software permite establecer atributos para metadatos espaciales (ver Figura 8). Incluso se puede definir un formato de metadatos a compartir en el entorno de la Salud Pública, que como se vio en el análisis es muy importante debido a la diversidad de maneras de realizar los estudios y obtener los datos.

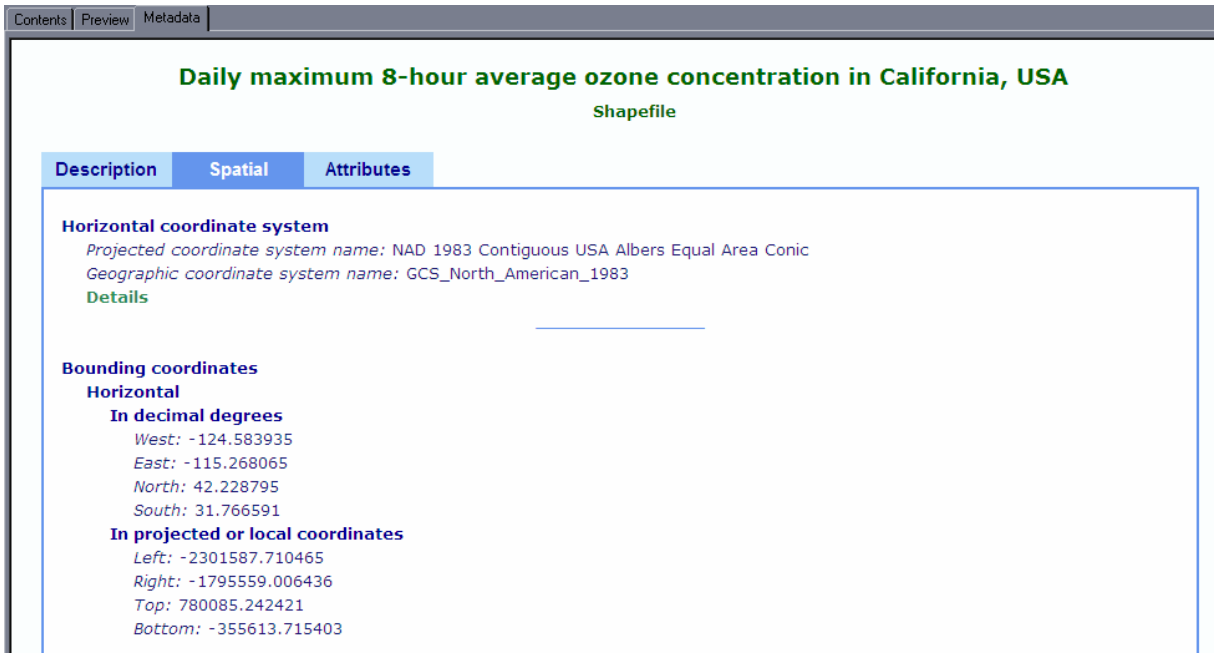


Figura 8 - La metadata puede tener que ver o no con datos espaciales.

En lo anterior se ven reflejados algunos requerimientos encontrados en la etapa de análisis (sobre todo en ENTRADAS) que un software de geoanálisis epidemiológico debería tener.

Luego de que los datos se ingresan (y sus metadatos se establecen) los mismos se pueden importar al ArcMap allí se utilizan o bien *toolbox* preexistentes en el programa o bien *toolbox* creados por nosotros.

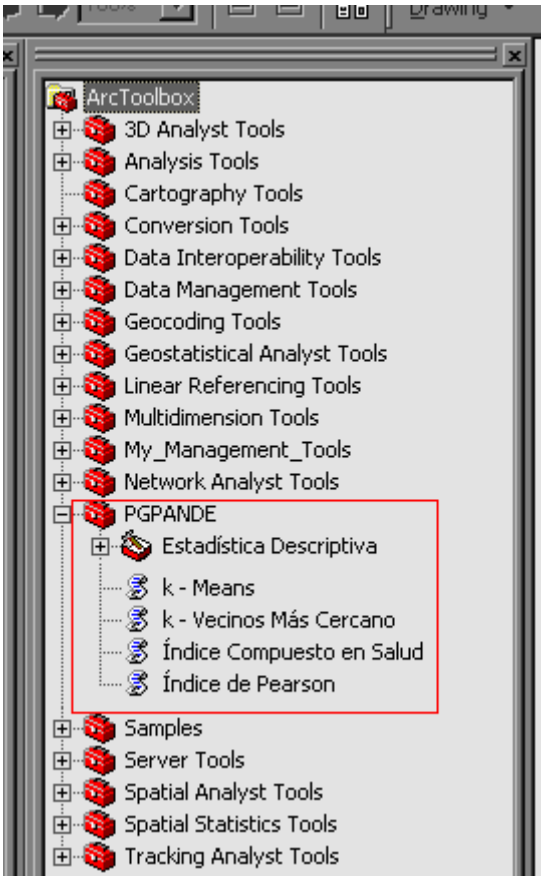


Figura 9 - En la figura se muestran algúnas *toolbox* del programa. Como destaque tenemos recuadrado en rojo algunas de las herramientas creadas para el proyecto.

En nuestro caso, y como ArcGIS carecía de herramientas específicas para análisis epidemiológico, desarrollamos herramientas para este fin utilizando Python y Java[20]. Además, creamos herramientas concretas para Minería de Datos.

Todos los datos pueden exportarse (luego del procesamiento que realiza cada herramienta en particular) a formatos de datos de uso estándar en SIG, como son los *shapefiles*[21]; Se pueden ingresar en una *geodatabase,* o bien se pueden exportar a un archivo de texto plano (como caso más primitivo).

Capítulo 4

Implementación

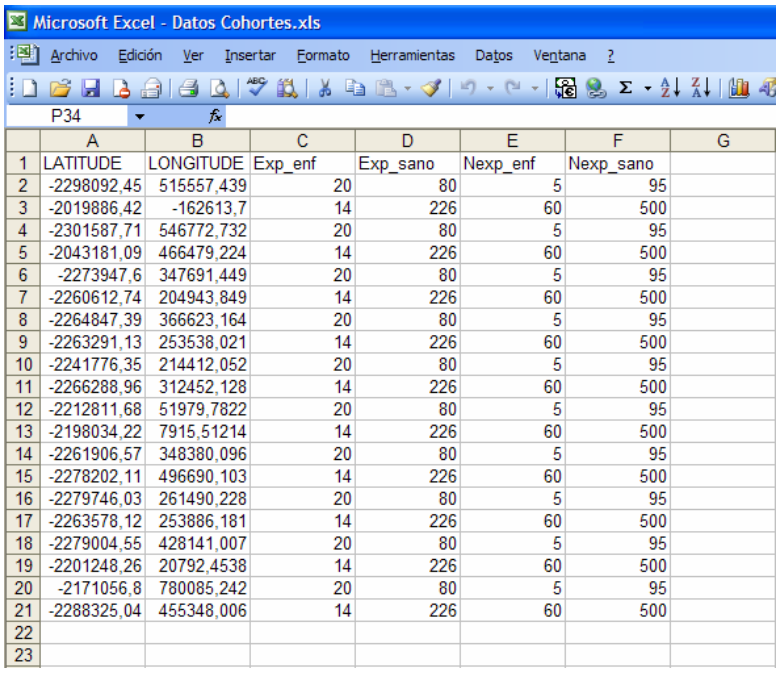
En lo siguiente explicaremos las herramientas implementadas y algunas que ya estaban presentes en el software pero que por su importancia en epidemiología se pensó que era necesario un estudio más profundo y detenido de las mismas, así como también su uso en el ArcMap. La mayoría de las herramientas implementadas fueron realizadas con Python y solo se usó Java para mostrar una ventana de resultados (ya que era más rápida su construcción).

Procedimientos específicos de análisis epidemiológicos

Estudio de Cohortes

Este estudio consiste básicamente en el seguimiento de uno o mas grupos de individuos sanos que presentan diferentes grados de exposición a un factor de riesgo en quienes se mide la aparición de la enfermedad en estudio.

La presente herramienta realiza un cohorte cerrada (todos los individuos fueron tomados en el mismo momento de tiempo) de dos grupos, uno expuesto al factor de riesgo y otro no. Utiliza un archivo de base de datos de ingreso (.dbf) para cada punto que se quiere realizar el cálculo el cual se genera exportando los datos cargados en una tabla Excel (Figura 10).


Microsoft Excel - Datos Cohortes.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

P34

	A	B	C	D	E	F	G
1	LATITUDE	LONGITUDE	Exp_enf	Exp_sano	Nexp_enf	Nexp_sano	
2	-2298092,45	515557,439	20	80	5	95	
3	-2019886,42	-162613,7	14	226	60	500	
4	-2301587,71	546772,732	20	80	5	95	
5	-2043181,09	466479,224	14	226	60	500	
6	-2273947,6	347691,449	20	80	5	95	
7	-2260612,74	204943,849	14	226	60	500	
8	-2264847,39	366623,164	20	80	5	95	
9	-2263291,13	253538,021	14	226	60	500	
10	-2241776,35	214412,052	20	80	5	95	
11	-2266288,96	312452,128	14	226	60	500	
12	-2212811,68	51979,7822	20	80	5	95	
13	-2198034,22	7915,51214	14	226	60	500	
14	-2261906,57	348380,096	20	80	5	95	
15	-2278202,11	496690,103	14	226	60	500	
16	-2279746,03	261490,228	20	80	5	95	
17	-2263578,12	253886,181	14	226	60	500	
18	-2279004,55	428141,007	20	80	5	95	
19	-2201248,26	20792,4538	14	226	60	500	
20	-2171056,8	780085,242	20	80	5	95	
21	-2288325,04	455348,006	14	226	60	500	
22							
23							

Figura 10 - Planilla con datos de ejemplo para el estudio de Cohortes.

La información que contiene para cada punto es su posición en el mapa (Longitud, Latitud), la cantidad de personas expuestas al factor de riesgo que enfermaron (Exp_enf), que no enfermaron (Exp_sano) y las personas no expuestas que enfermaron (Nexo_enf) y las que no (Nexo_sano). Finalmente retorna un conjunto de índices útiles para la toma de decisiones, además de un campo que indica si la variable que se quiere testear como posible causa de la enfermedad lo es en realidad o no.

Los índices generados son:

- Tasa de incidencia de sujetos expuestos (Ti_exp) y no expuestos (Ti_noexp)
- La magnitud de riesgo asociado a la exposición o Riesgo Relativo (RR)
- Limite superior (RConf_inf) e inferior (RConf_sup) del intervalo de confianza del RR
- La diferencia aritmética entre incidencia de la enfermedad en expuestos y no expuestos o Riesgo atribuible (RA)
- El cociente entre RA y la incidencia de la enfermedad en expuestos o riesgo atribuible porcentual (RA_porc)
- La diferencia aritmética entre la incidencia de la enfermedad en la población general y la población no expuesta al factor de riesgo o riesgo atribuible poblacional (RAP)
- El cociente entre el RAP y la incidencia de la enfermedad en la población total por 100 o riesgo atribuible poblacional porcentual (RAP_porc)
- El resultado utilizando los índices anteriores (Asoc_RR), que es 'Riesgo significativo' en caso que RConf_inf >1 y es 'Protección significativa' si RConf_sup < 1

Para poder usar el procedimiento se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta "Cohortes"
3. Se abrirá el cuadro de la Figura 11.
    a. Seleccione el archivo de base de datos (.dbf) a utilizar
    b. Seleccione el archivo con la referencia espacial a usar
    c. Seleccione el nombre de la nueva capa que contendrá los valores resultado.
    d. Seleccione el botón OK

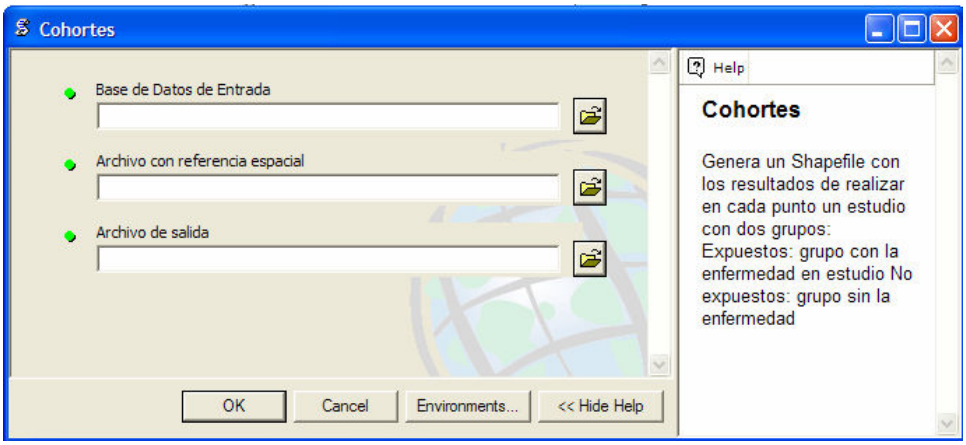


Figura 11 – Cuadro de la herramienta de Cohortes

Luego genera una capa para cada punto ingresado con los resultados, permitiendo colorear los puntos donde da positivo y donde no (Figura 12).

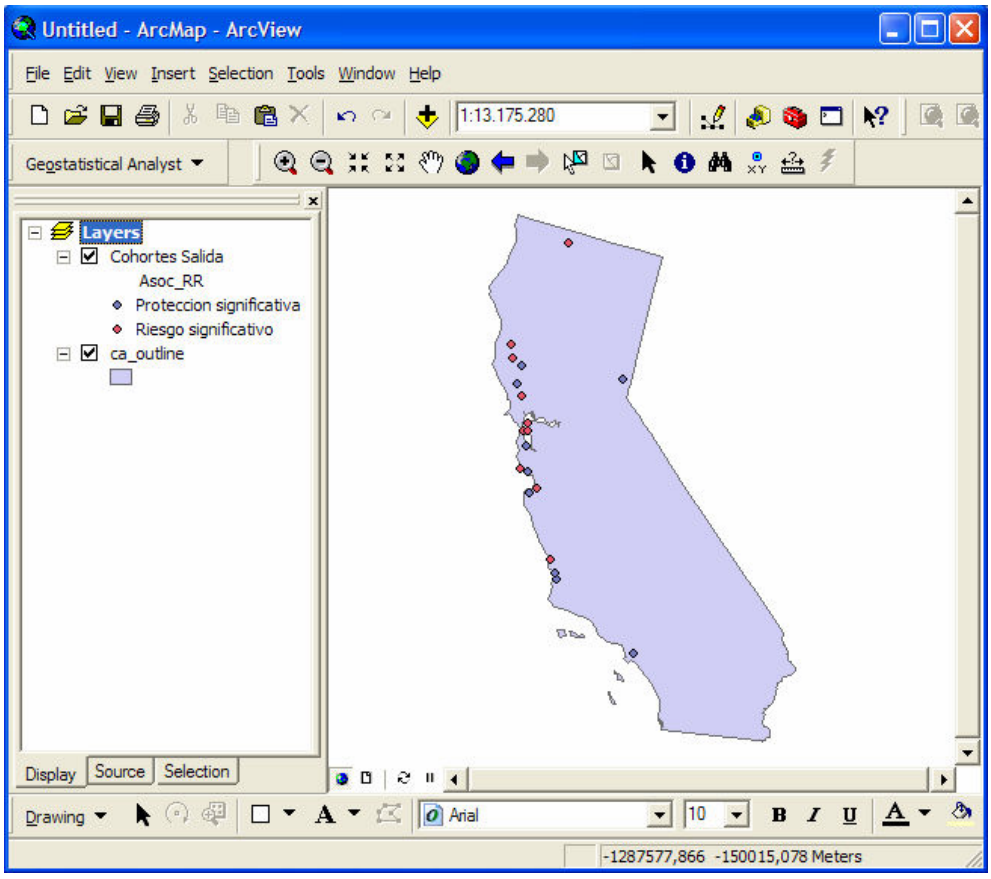
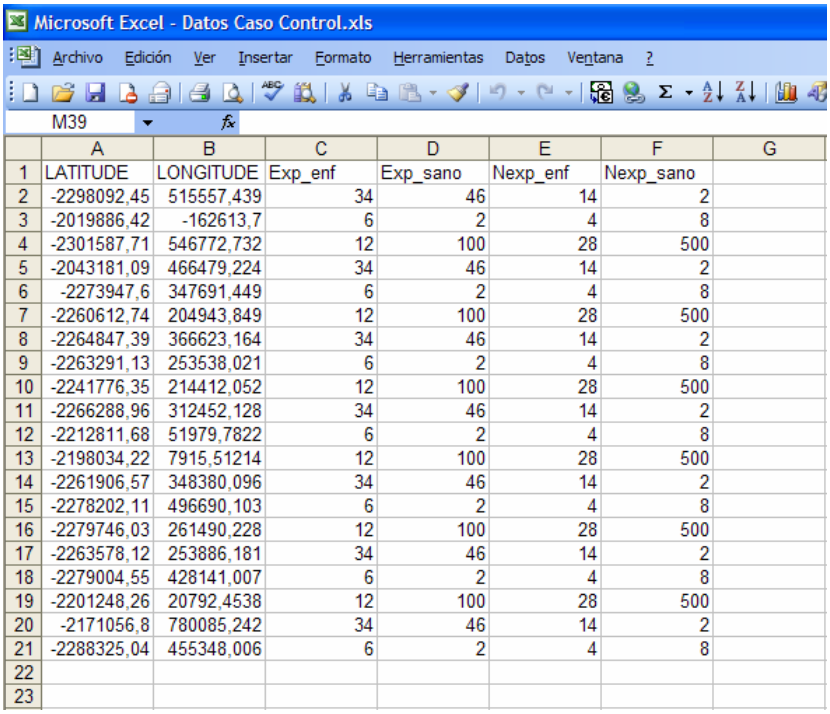


Figura 12 - Resultado de realizar el estudio y representarlo en el mapa.

Estudio Caso Control

A partir de un conjunto de puntos tomados de una base de datos (.dbf) que es generada a partir de una planilla Excel (Figura 13), los cuales contienen datos de un grupo "casos" (formando por un conjunto de sujetos que presentan la enfermedad o el caso de estudio que se quiere poner a prueba como hipótesis) y un grupo "control" de sujetos con las mismas características excepto que no presentan el caso a estudiar, la herramienta genera una capa con los resultados obtenidos de los datos de ingreso. También se necesita ingresar un archivo con la referencia espacial que se desea usar para generar la capa resultado.

A partir de estos define que tipo de asociación hay entre la exposición al factor causante de la enfermedad y el resultado final (Riesgo significativo o Protección significativa). Esta asociación se puede usar como clave para representar gráficamente los resultados y sacar conclusiones para aceptar o refutar la hipótesis original (Figura 15).


Microsoft Excel - Datos Caso Control.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

M39

	A	B	C	D	E	F	G
1	LATITUDE	LONGITUDE	Exp_enf	Exp_sano	Nexp_enf	Nexp_sano	
2	-2298092,45	515557,439	34	46	14	2	
3	-2019886,42	-162613,7	6	2	4	8	
4	-2301587,71	546772,732	12	100	28	500	
5	-2043181,09	466479,224	34	46	14	2	
6	-2273947,6	347691,449	6	2	4	8	
7	-2260612,74	204943,849	12	100	28	500	
8	-2264847,39	366623,164	34	46	14	2	
9	-2263291,13	253538,021	6	2	4	8	
10	-2241776,35	214412,052	12	100	28	500	
11	-2266288,96	312452,128	34	46	14	2	
12	-2212811,68	51979,7822	6	2	4	8	
13	-2198034,22	7915,51214	12	100	28	500	
14	-2261906,57	348380,096	34	46	14	2	
15	-2278202,11	496690,103	6	2	4	8	
16	-2279746,03	261490,228	12	100	28	500	
17	-2263578,12	253886,181	34	46	14	2	
18	-2279004,55	428141,007	6	2	4	8	
19	-2201248,26	20792,4538	12	100	28	500	
20	-2171056,8	780085,242	34	46	14	2	
21	-2288325,04	455348,006	6	2	4	8	
22							
23							

Figura 13 - Planilla con datos de ejemplo para el estudio de Caso control.

Los índices generados por la herramienta son:

- Tasa de exposición entre los sujetos casos (Texp_casos) y los controles (Texp_ctrls)
- La división entre la chance de tener la enfermedad y la de no tenerla u odds ratio (OddsRatio)
- Limite superior (RConf_inf) e inferior (RConf_sup) del intervalo de confianza del OddsRatio
- El porcentaje de protección sobre la población estudiada (Porc_prot)
- El resultado utilizando los índices anteriores (Asoc_OR), que es 'Riesgo significativo' en caso que RConf_inf >1 y es 'Protección significativa' si RConf_sup < 1

Para poder usar el procedimiento se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta “Caso Control”
3. Se abrirá el cuadro de la Figura 14.
    a. Seleccione el archivo de base de datos (.dbf) a utilizar
    b. Seleccione el archivo con la referencia espacial a usar
    c. Seleccione el nombre de la nueva capa que contendrá los valores resultado.
    d. Seleccione el botón OK

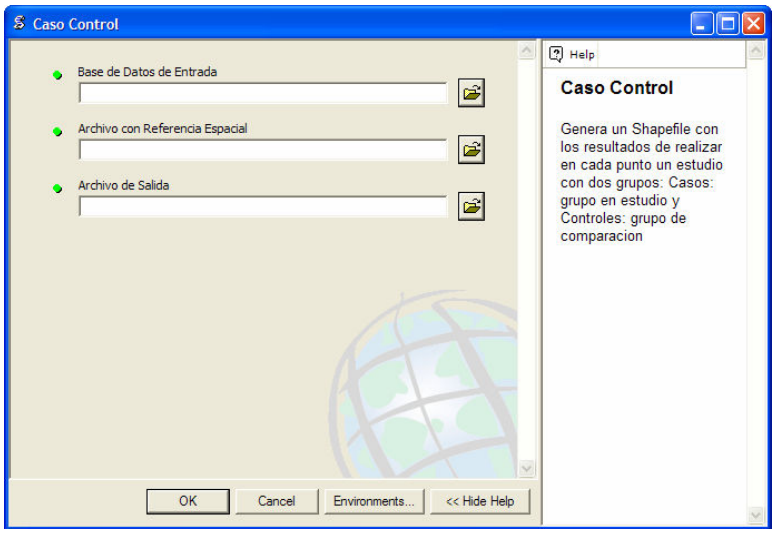
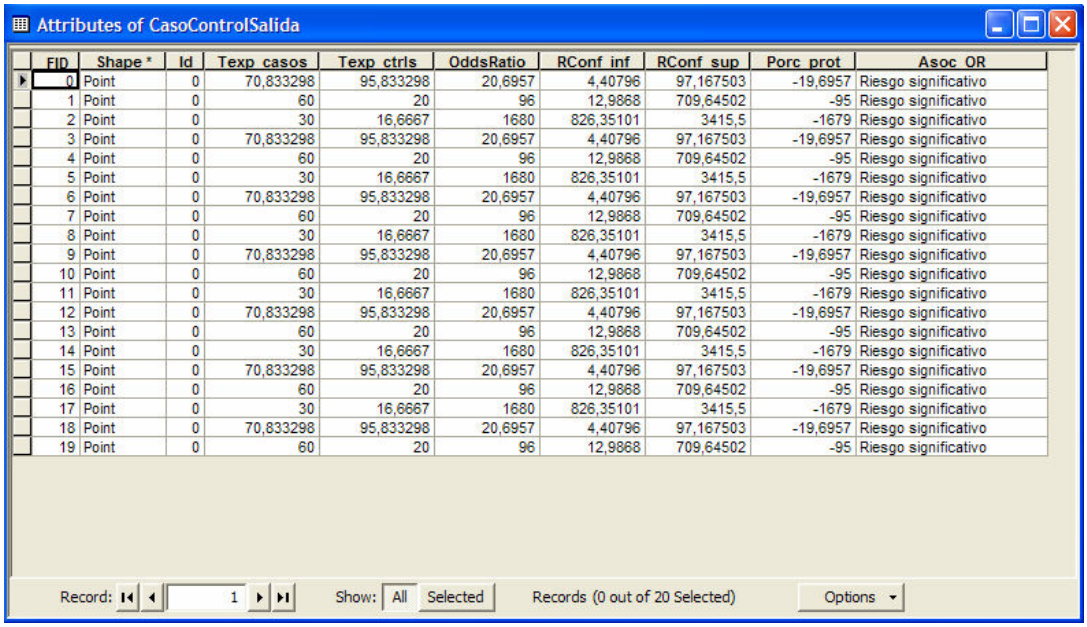


Figura 14 - Cuadro de la herramienta caso control



FID	Shape *	Id	Texp_casos	Texp_ctrls	OddsRatio	RConf_inf	RConf_sup	Porc_prot	Asoc_OR
0	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
1	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
2	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
3	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
4	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
5	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
6	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
7	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
8	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
9	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
10	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
11	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
12	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
13	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
14	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
15	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
16	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo
17	Point	0	30	16,6667	1680	826,35101	3415,5	-1679	Riesgo significativo
18	Point	0	70,833298	95,833298	20,6957	4,40796	97,167503	-19,6957	Riesgo significativo
19	Point	0	60	20	96	12,9868	709,64502	-95	Riesgo significativo

Figura 15 - Resultado de utilizar la herramienta.

Procedimientos generales de análisis y estadística de datos

Estadística Descriptiva y Distribución de Frecuencias

La Estadística Descriptiva es una función sencilla del análisis exploratorio de datos. Permite calcular un conjunto de medidas de tendencia central y dispersión de valores observados de una variable, dato o indicador de salud. La Distribución de Frecuencia permitirá cuantificar la frecuencia de valores de una variable, es decir, la cantidad de veces que se repite el valor observado o registrado de una variable objeto del análisis.

Para poder usar el procedimiento de análisis (que viene dentro del ArcMap) se deberá realizar lo siguiente (para ArcMap):

1. Seleccione la capa y las *features* dentro de esa capa a la que quiere hacerle el análisis.

2. Ir al menú "*Selection*", luego seleccionar "*Statistics…*".

3. En la Figura 16 se muestran las estadísticas de las *features* seleccionadas (dentro del cuadrado azul) y la gráfica de distribución de frecuencias de las mismas (dentro del cuadro rojo).

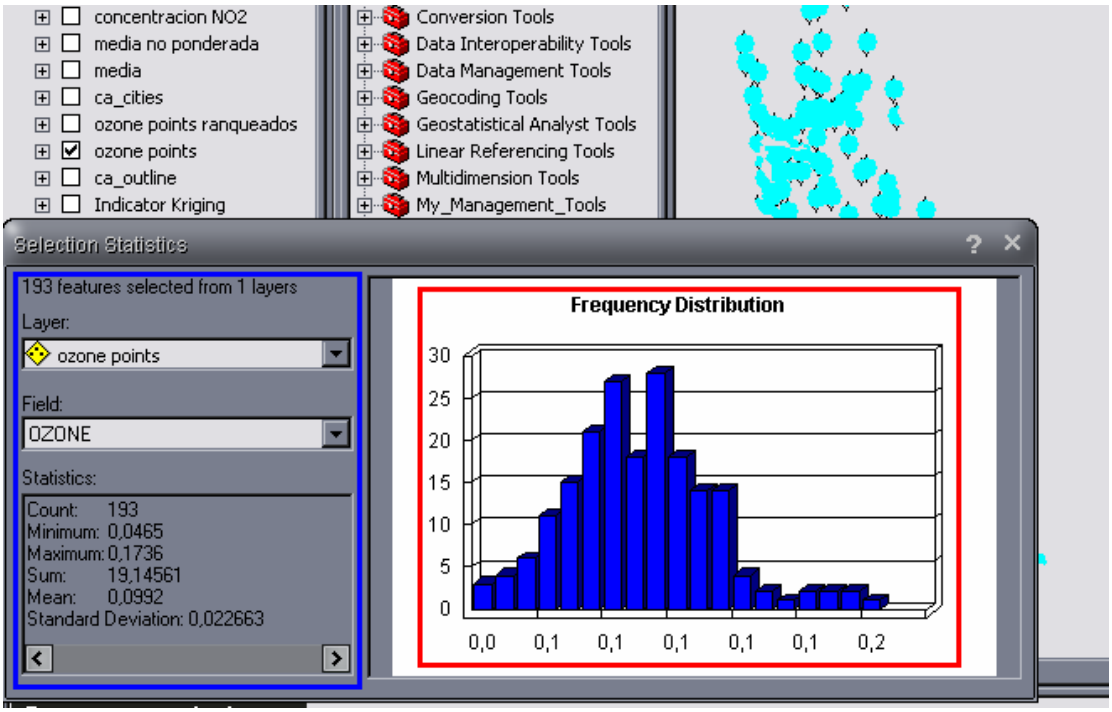


Figura 16 - Estadística Descriptiva y Distribución de Frecuencias aplicadas a datos seleccionados.

Análisis de Correlación

Es un método estadístico que permite medir el nivel de correlación entre dos ó más variables, presentando los resultados en una matriz de correlación. Este método calcula el Coeficiente de Correlación de Pearson[22].

El Coeficiente de Correlación mide el grado y la dirección de "co-relación" lineal entre las variables. El coeficiente de correlación elevado al cuadrado describe la proporción de co-variación entre las variables.

Para poder usar el procedimiento de análisis se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta "Índice de Pearson"
3. Se abrirá el cuadro de la Figura 17.
    a. Seleccione la capa a la que se le quiere realizar el cálculo
    b. Seleccione la primer variable a la que quiere correlacionar
    c. Seleccione la segunda variable a correlacionar
    d. Seleccione el nivel de significancia del cálculo del índice (por defecto se calculará con una significancia del 95%)

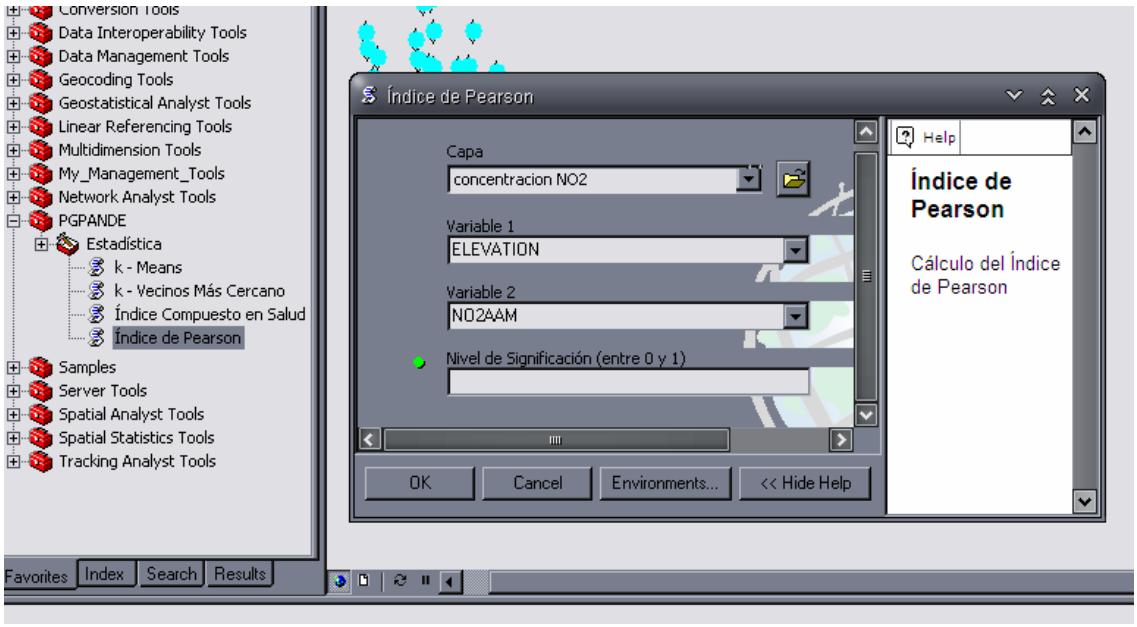


Figura 17 - Análisis de Correlación entre variables. Cuadro de entrada de datos.

El resultado se muestra en la Figura 18. El resultado es un índice Real con valor entre 1 y -1 con 0 significando que no están correlacionados, -1 que tienen correlación inversa y 1 correlación directa.

Se muestra además el test de hipótesis para probar la significancia ingresada y su respectivo intervalo de confianza para el valor de pearson hallado.

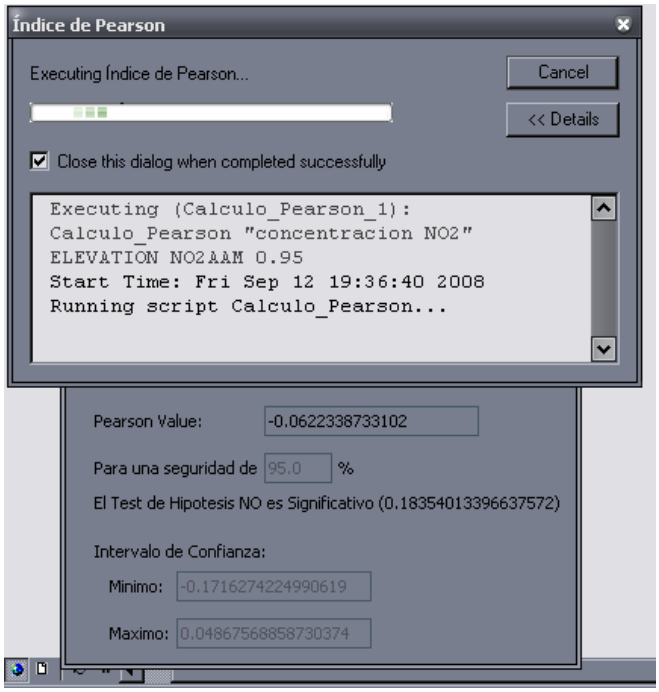


Figura 18 - Análisis de Correlación entre variables. Cuadro de salida de resultados.

Procedimientos generales para análisis y estadística epidemiológica

Estandarización de Tasas

Se generaron dos herramientas para realizar la estandarización de población de datos muy dispares o incompatibles entre si a través del método directo o indirecto.
En ambos casos se utiliza un archivo de base de datos (.dbf) con los valores en bruto, en el caso directo se necesita además una base con un conjunto de población estándar (Figura 19) y en el caso indirecto una base con tasas estándar (Figura 20). Estas bases estándar pueden ser obtenidas en universidades o sitios especializados.

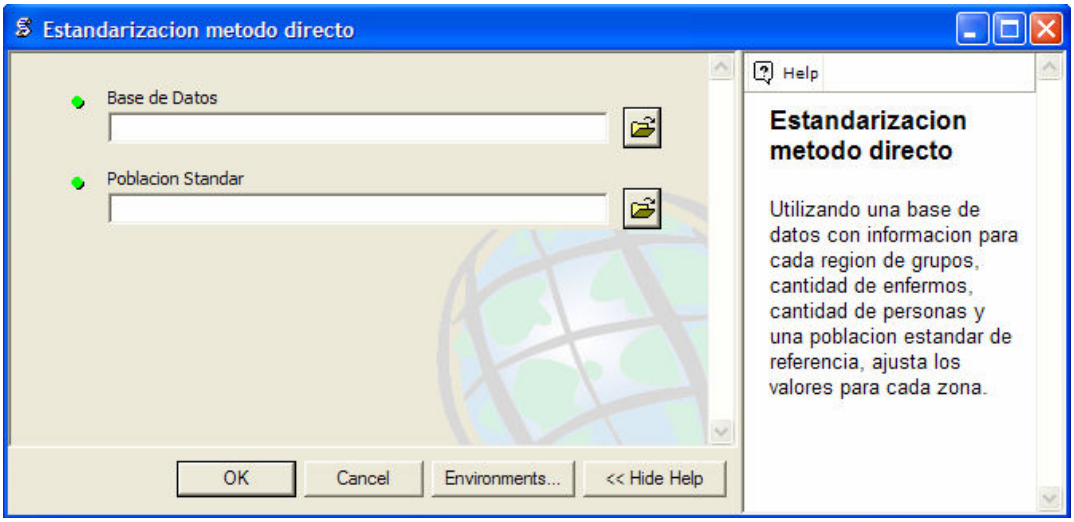


Figura 19 - Ingreso de datos de la herramienta para el caso directo.

El resultado se almacena en otra base de datos (.dbf) lo cual permite generar una capa *shapefile* o ser utilizada como punto de partida de otras herramientas.

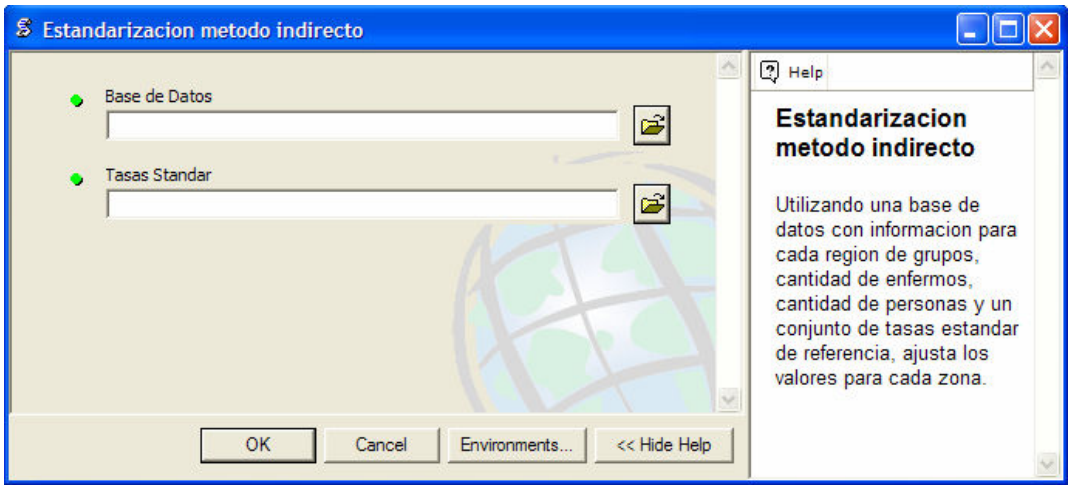


Figura 20 - Ingreso de datos de la herramienta para el caso indirecto.

Identificación de Áreas Críticas

El procedimiento de Identificación de Áreas Críticas permite definir un criterio utilizando medidas epidemiológicas e indicadores previamente calculados. Como resultado, se seleccionan en el mapa las *features* que cumplen la condición previamente establecida.

Este procedimiento es útil en los procesos de asignación de recursos, planificación de acciones de control y prevención, y definición áreas de mayor riesgo ante la aparición de eventos que hagan peligrar la salud de la población en estudio.

Para poder usar el procedimiento de análisis (que viene dentro del ArcMap) se deberá realizar lo siguiente (para ArcMap):

1. Haga clic en el menú "Selection" y seleccione "*Select by Attributes…*'
2. En la ventana de la Figura 21 seleccione: la capa y luego genere con las columnas y los valores lógicos una sentencia lógica como la que muestra la figura.
3. Finalmente presione sobre el botón "OK" o "Apply"

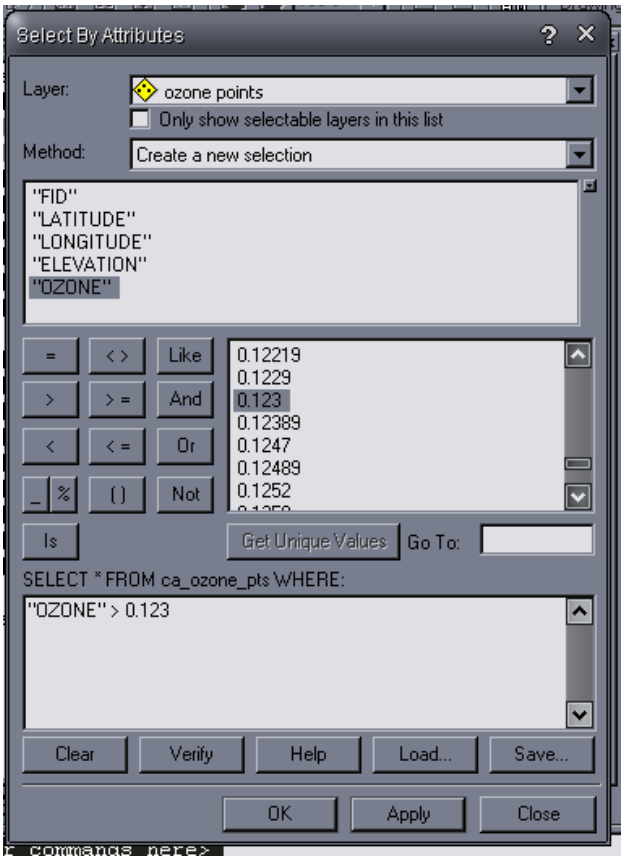


Figura 21 - La figura muestra un ejemplo de expresión lógica sencilla para valores de ozono > 0,123

Como resultado obtenemos las figuras geográficas que cumplen la condición lógica (mostrados en celeste en la Figura 22).

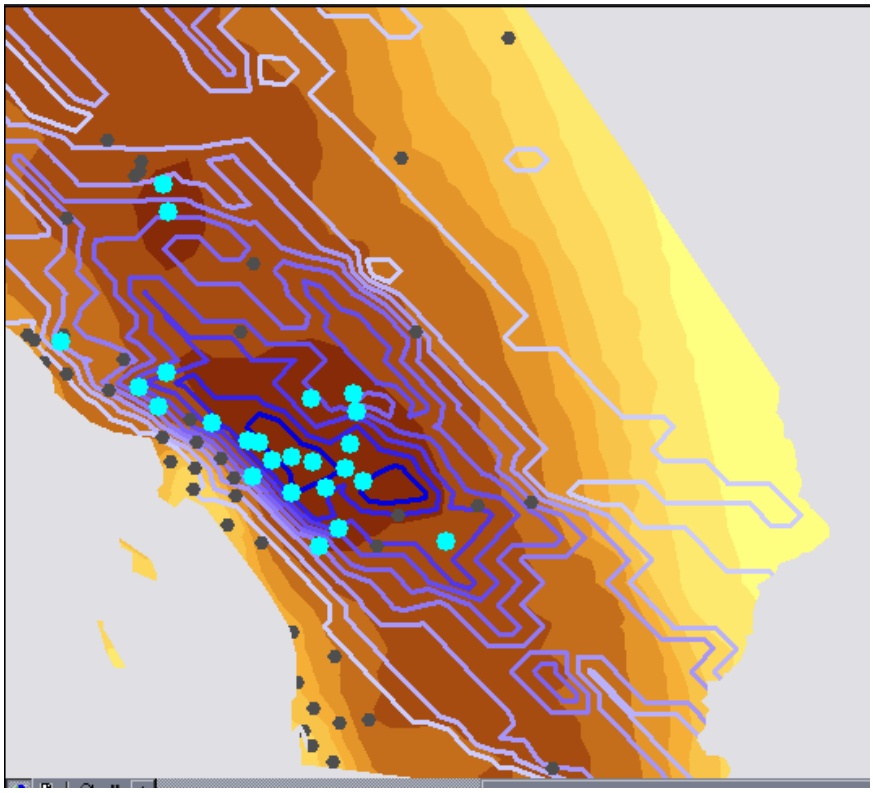


Figura 22 - Resultado de la selección de áreas criticas

El resultado del procedimiento de selección de áreas críticas. Según la Figura 22 en celeste están los elementos que cumplen la condición y en verde oscuro los que no.

Cálculo de un Índice Compuesto en Salud

El uso del procedimiento de ICS es muy útil para análisis de situación de salud en un área o unidad geográfica a partir de varios indicadores que se poseen de salud y se desea agrupar o unir estos indicadores y crear un índice que sintetice y refleje el comportamiento de todos los indicadores que participan en el análisis.

Este índice está basado en *Zscores*. Para cada indicador se calcula su media y su desviación estándar y a partir de aquí se calcula su valor Z. El Indice Compuesto en Salud será la suma de todos los Z de cada indicado r, con

$$Z = \frac{(I - m)}{d},$$

Donde I es el valor del indicador, m es la media y d su desviación estándar.

Generando así el Índice Compuesto en Salud:

$$ICS = a_1.Z_1 + ... + a_n.Z_n ,$$

Donde a_i es el peso que tiene el indicador Z_i, sabiendo que $\sum_i |a_i| = 1$

Para poder usar el procedimiento de análisis se deberá realizar lo siguiente (para ArcMap):

4. Expanda las *toolbox* de PGPANDE.
5. Seleccione la herramienta "Índice Compuesto en Salud"
6. Se abrirá el cuadro de la Figura 23.
    a. Seleccione la capa a la que se le quiere realizar el cálculo
    b. Seleccione las variables de la capa que sean de peso positivo
    c. Seleccione los pesos respectivos de las variables de la capa que sean de peso positivo
    d. Seleccione las variables de la capa que sean de peso negativo
    e. Seleccione los pesos respectivos de las variables de la capa que sean de peso negativo.
    f. Seleccione el nombre de la nueva capa que contendrá los valores de ICS.
    g. Seleccione el botón OK

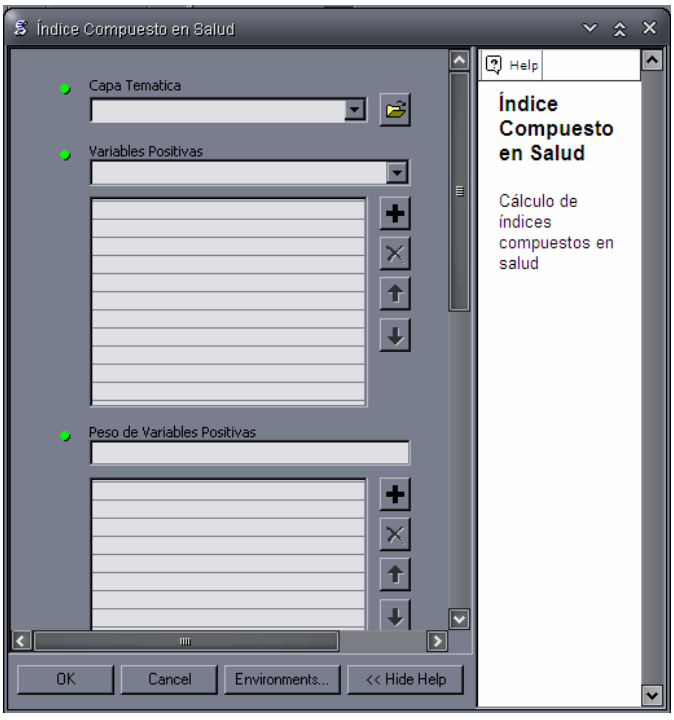


Figura 23 - Ventana de cálculo del Índice Compuesto en Salud

El resultado se muestra en la Figura 24. Se genera una capa de puntos con

los centroides de las *features* de la capa de entrada. La razón de esto es que tener puntos nos facilita a la hora de generar mapas de tendencia, raster, etc.

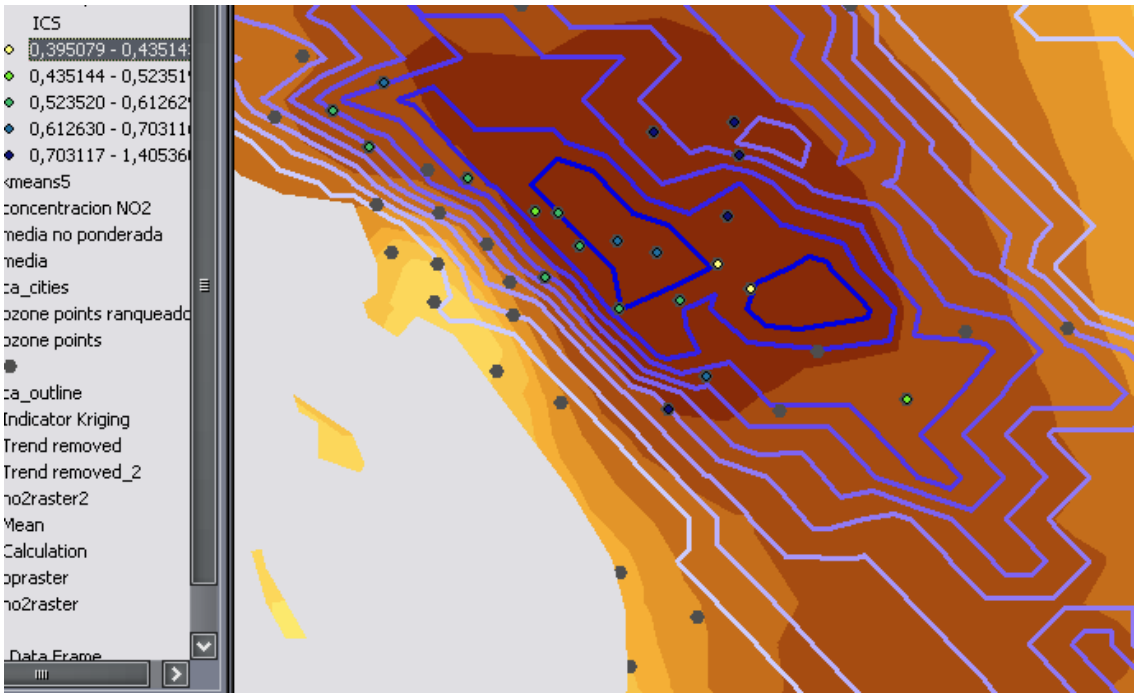


Figura 24 - Resultado del cálculo de ICS para puntos de Ozono. Se pueden ver a la izquierda los colores que indican distintos niveles del índice (cuando más oscuro mayor es el valor del índice).

Procedimientos para análisis espacial

Suavizador espacial

Esta herramienta tiene como entrada una capa de puntos y una columna de la misma, permitiendo suaviza los valores de dicha columna (Figura 25).

Primero forma un mapa *raster* haciendo una interpolación de los puntos brindados y es a este mapa que se le suavizan sus valores tomando para cada punto un entorno y promediando sus valores (Se muestra el modelo de la herramienta en la Figura 26). Esto genera una capa *raster* como resultado.

Para poder usar el procedimiento se deberá realizar lo siguiente (para ArcMap):

1. Expanda las toolbox de PGPANDE.

2. Seleccione la herramienta “Suavizador Espacial”
3. Se abrirá el cuadro de la (agregar numero de la Figura 25).
    a. Seleccione el archivo de capa de puntos a utilizar.
    b. Seleccione la variable a suavizar.
    c. Seleccione el nombre de la nueva capa que contendrá los valores resultado.
    d. Seleccione el botón OK

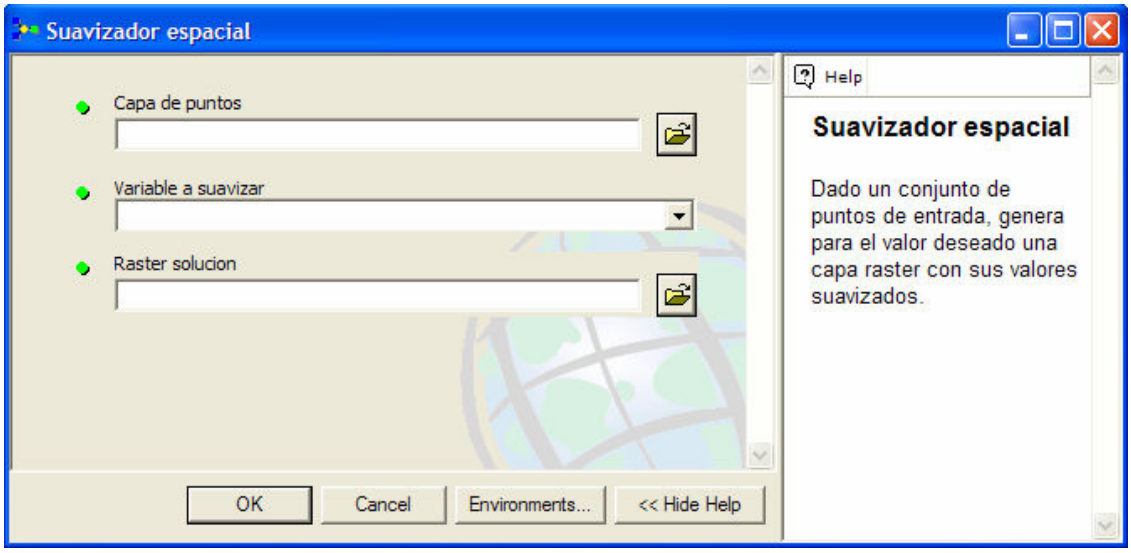


Figura 25 - Ingreso de datos de la herramienta.

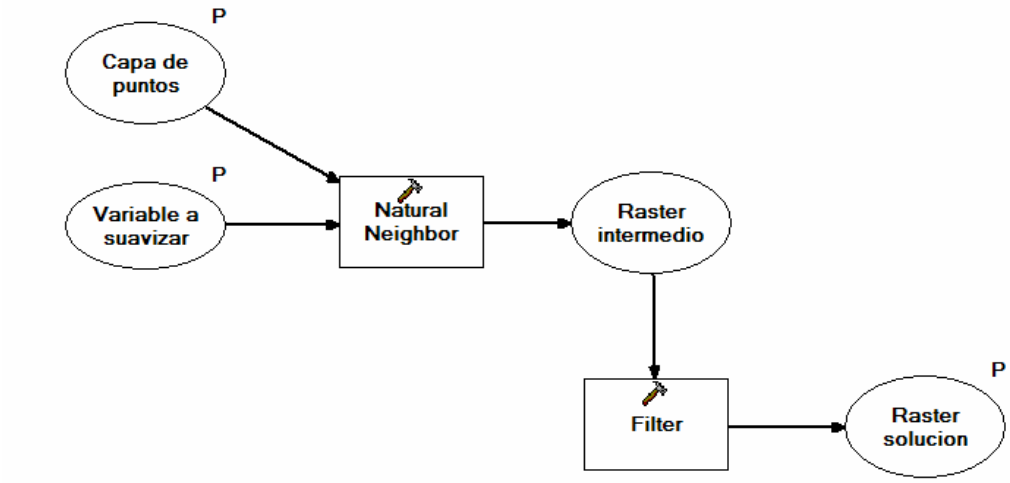


Figura 26 - Modelo para realizar el suavizado de los valores.

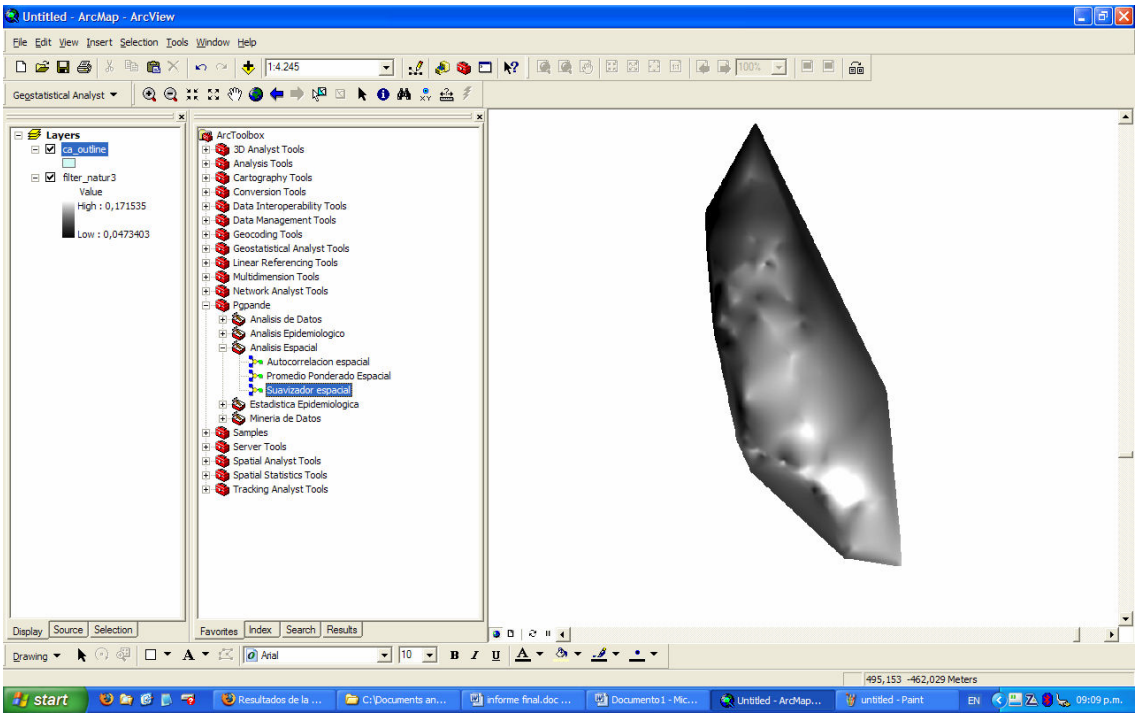


Figura 27 - Resultado de la herramienta suavizador espacial sobre datos de ozono en California

Valor promedio ponderado espacial

A partir de una capa de puntos, la variable deseada para generar el promedio ponderado y el tipo y tamaño del vecindario que se desea utilizar para tomar como entorno de los valores a promediar, genera una capa *raster* mostrando gráficamente los valores promedio (Figura 28).

Para poder usar el procedimiento se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta “Promedio Ponderado Espacial”
3. Se abrirá el cuadro de la Figura 28.
    a. Seleccione el archivo de capa de puntos a utilizar.
    b. Seleccione la variable a la que se quiere realizar el promedio.
    c. Seleccione el tipo de vecindario a tomar en cuenta por la herramienta (circular, irregular, rectangular, etc.).
    d. Seleccione la distancia entre valores y la unidad que se va a usar para los cálculos.
    e. Seleccione el nombre de la nueva capa que contendrá los valores resultado.

f. Seleccione el botón OK

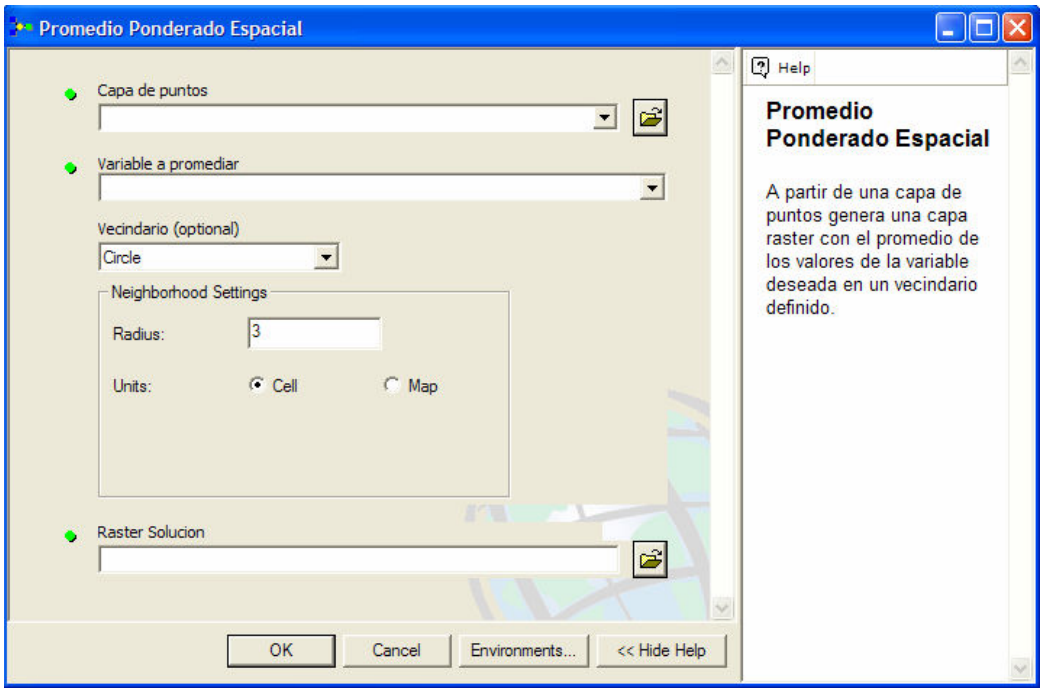


Figura 28 - Entrada de datos de la herramienta.

La herramienta internamente consta de dos pasos, el primero realiza una interpolación a partir de los puntos de ingreso y genera una capa *raster* intermedia. El segundo paso toma la capa intermedia y se le aplica una función que calcula localmente los valores promedio. La forma y tamaño del entorno local puede ser modificado por el usuario (Figura 29).

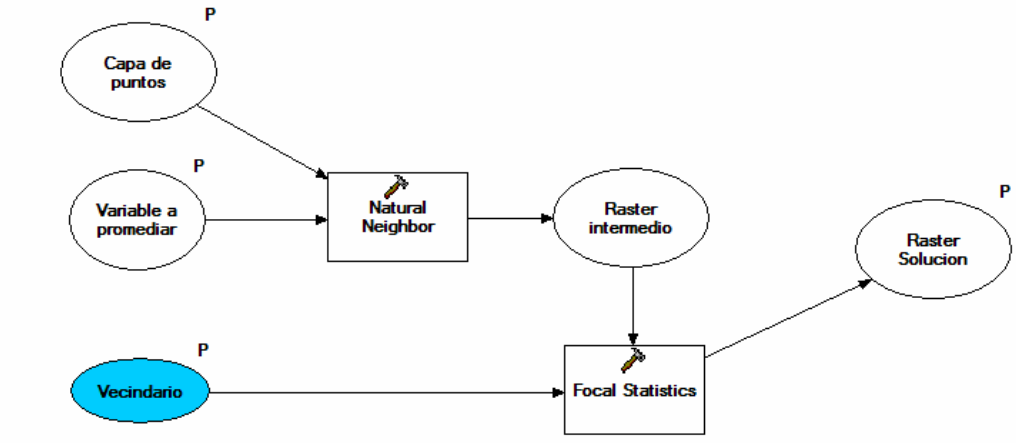


Figura 29 - Modelo para realizar el promedio ponderado espacial.

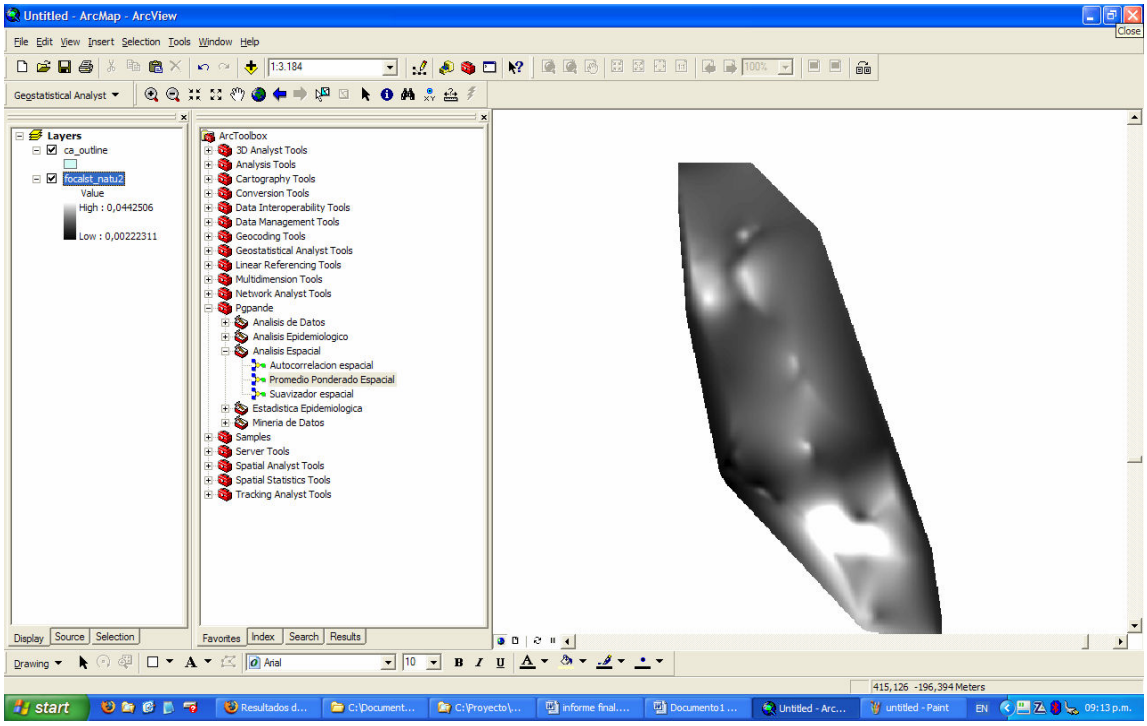


Figura 30 - Resultado de utilizar promedio ponderado espacial con datos de NO2 en California

Índices de autocorrelación espacial

Permite calcular la matriz de correlación tomando en cuenta la posición de los objetos que contiene una capa de *features* y el valor del atributo deseado. Debe estar marcada la opción de mostrar gráficamente para que aparezca la ventana solución, también se debe elegir el método a utilizar para el cálculo de las distancias entre los objetos, que puede ser Euclidiano o Manhattan (Figura 31).

Para poder usar el procedimiento se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta “Autocorrelación Espacial”
3. Se abrirá el cuadro de la Figura 31.
    a. Seleccione el archivo de capa de puntos a utilizar.
    b. Seleccione la variable que se quiere ver si sus datos están dispersos o no.

c. Seleccione la opción de “Mostrar la salida gráficamente”.
d. Seleccione el tipo de distancia a usar.
e. Seleccione el botón OK

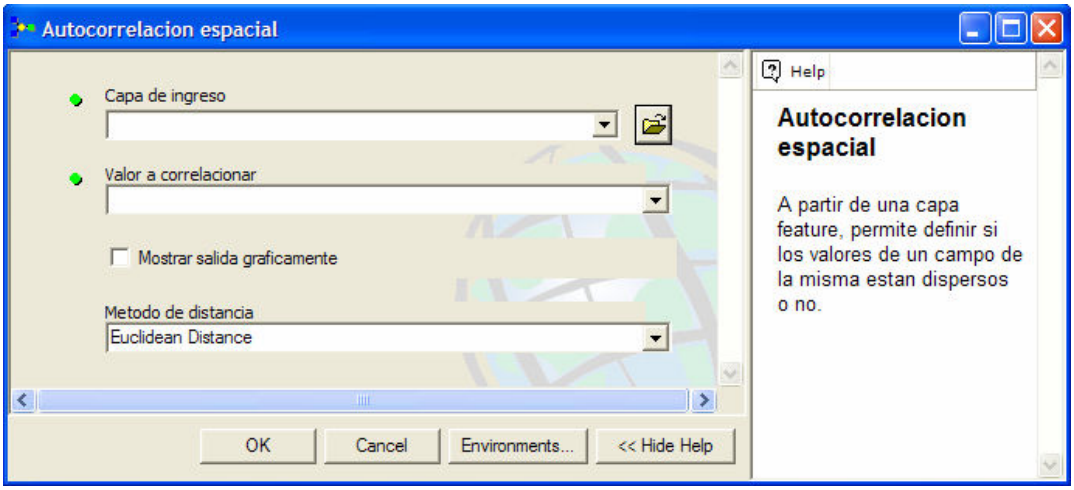


Figura 31 - Ventana de entrada de datos.

Según el valor de resultado especifica si lo datos seleccionados están dispersos en el mapa o agrupados (*clustered*), esto permitiría en un futuro generar alertas automáticas de epidemia en el caso de calcular la agrupación de casos de ciertas enfermedades.

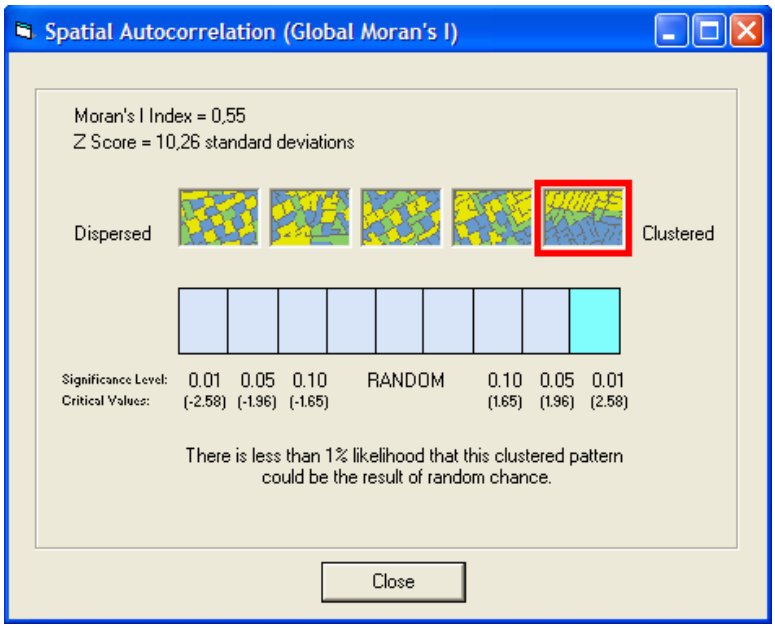


Figura 32 - Resultado de la herramienta con datos de NO2 en California

Procedimientos para descubrimiento de información

Árboles de decisión

Como ya se ha visto en el estado del arte, árboles de decisión es un método para clasificar información y descubrir relaciones entre datos que en un principio parecen desconectados. El algoritmo utilizado se denomina ID3 y es un clásico dentro de las heurísticas existentes. Genera una buena solución para el problema planteado utilizando el concepto de entropía para calcular cual es el mejor atributo por el cual subdividir los datos en subconjuntos. Para la presentación en pantalla de los resultados se utilizaron librerías contenidas en un paquete *open source* de aplicaciones GUI implementadas en python llamado wxPython[23].

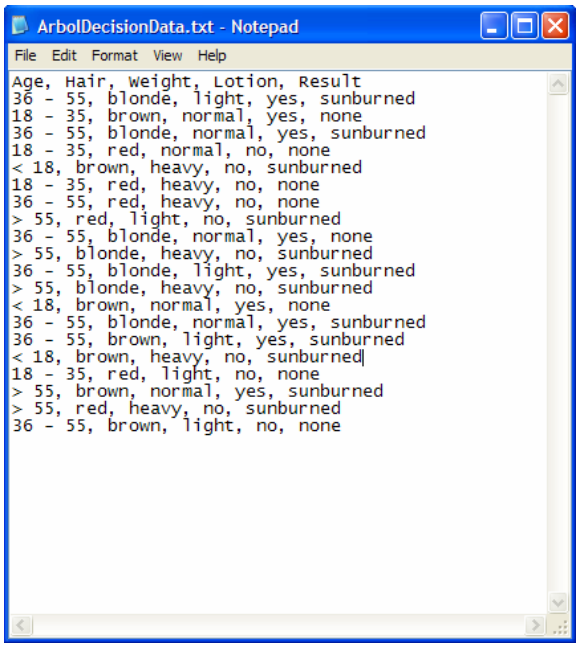


Figura 33 - Archivo de ingreso con los datos de ejemplo.

A partir de un archivo de texto plano (Figura 33) con datos obtenidos por censos u otros medios, el algoritmo los procesa y el resultado obtenido se pueden observar en la Figura 34. El ejemplo de la figura fue generado por nosotros con datos inventados pero si estos fueran ciertos se podrían deducir relaciones que en un comienzo no estaban visibles, como por ejemplo que los niños sin protector están expuestos a quemaduras de sol y que los ancianos son los mas sensibles al mismo.

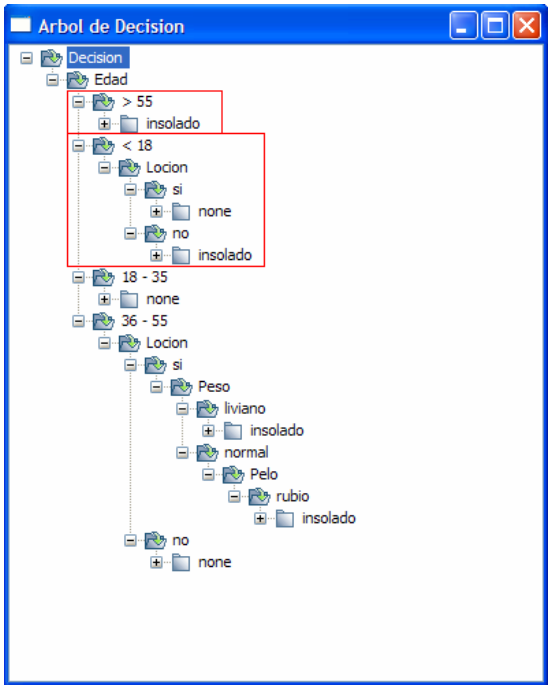


Figura 34 - Árbol de decisión resultante del ejemplo anterior.

K-Means

El procedimiento realiza una clusterización por *k-means* y separa en grupos de similaridad a la capa de entrada. Para esto se usa una función de distancia que depende del tipo de datos a comparar.

Para poder usar el procedimiento de análisis se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta “k-Means”
3. Se abrirá el cuadro de la Figura 35.
    a. Seleccione la cantidad de clusters en que quiere separar.
    b. Seleccione la capa a la que se le quiere realizar el cálculo
    c. Seleccione las variables independientes de la capa que quiere que influencien en la clusterización.
    d. Seleccione el nombre de la nueva capa que contendrá los valores de k-menas.
    e. Seleccione OK

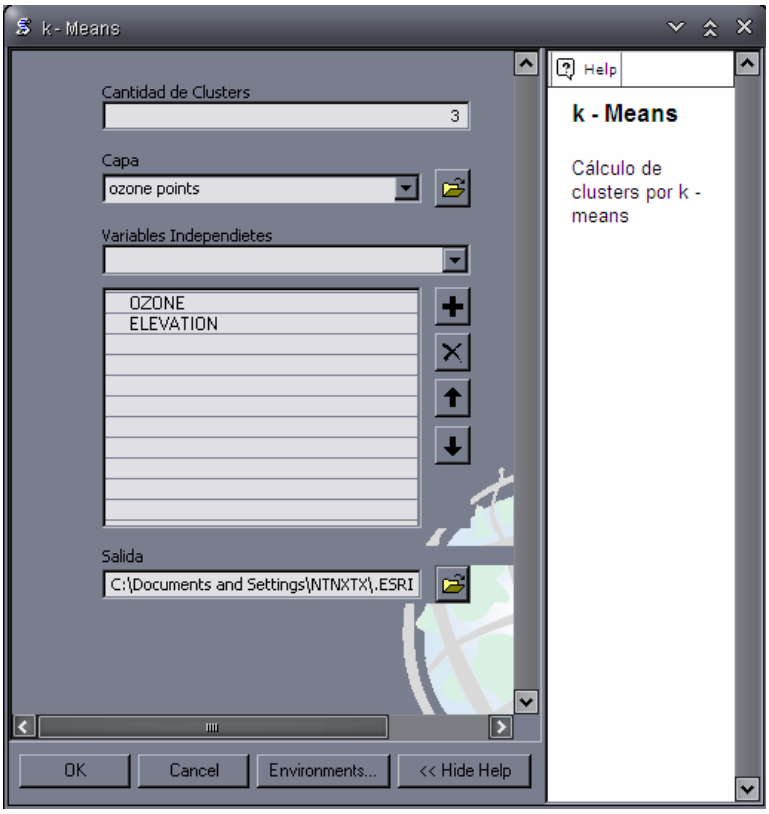


Figura 35 - Ventana de Ingreso de variables para el cálculo de k-means.

El resultado es una capa de puntos que contienen los centroides de la capa de entrada y el valor del cluster al cual pertenece (Figura 36).

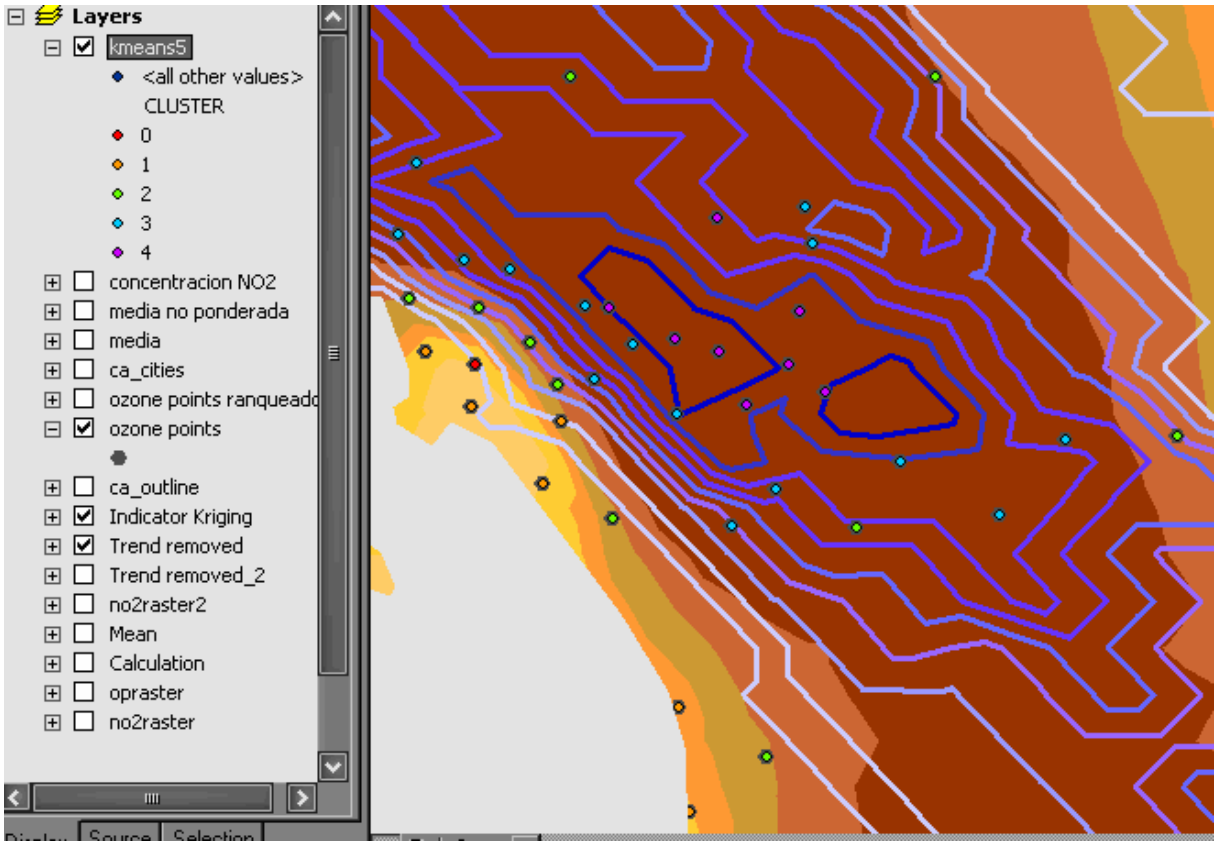


Figura 36 - A la izquierda se muestra el resultado de k-means con 5 cluster y sus respectivos colores reflejados en el mapa de la derecha.

K-vecinos más cercanos

El procedimiento de K-Vecinos Más Cercano busca cuales son los vecinos más cercanos a un elemento con respecto a un conjunto de variables independientes, utilizando una "distancia" para medir cuan cercanos están los puntos de referencia al nuevo elemento y devuelve una capa de puntos con los centroides de los vecinos más cercanos.

Para poder usar el procedimiento de análisis se deberá realizar lo siguiente (para ArcMap):

1. Expanda las *toolbox* de PGPANDE.
2. Seleccione la herramienta "k - Vecinos Más Cercanos" (Figura 37).
3. Seleccione la cantidad de Vecinos Más Cercanos.
4. Seleccione las variables independientes usadas para generar el cálculo de *knn*
5. Seleccione las variables dependientes usada para estimación.
6. Seleccione tantos nuevos valores como variables (y en el mismo orden)

agregó en el punto 4.

7. Ingrese la capa de puntos de Vecinos Más Cercanos resultado.
8. Presione OK para realizar el cálculo.

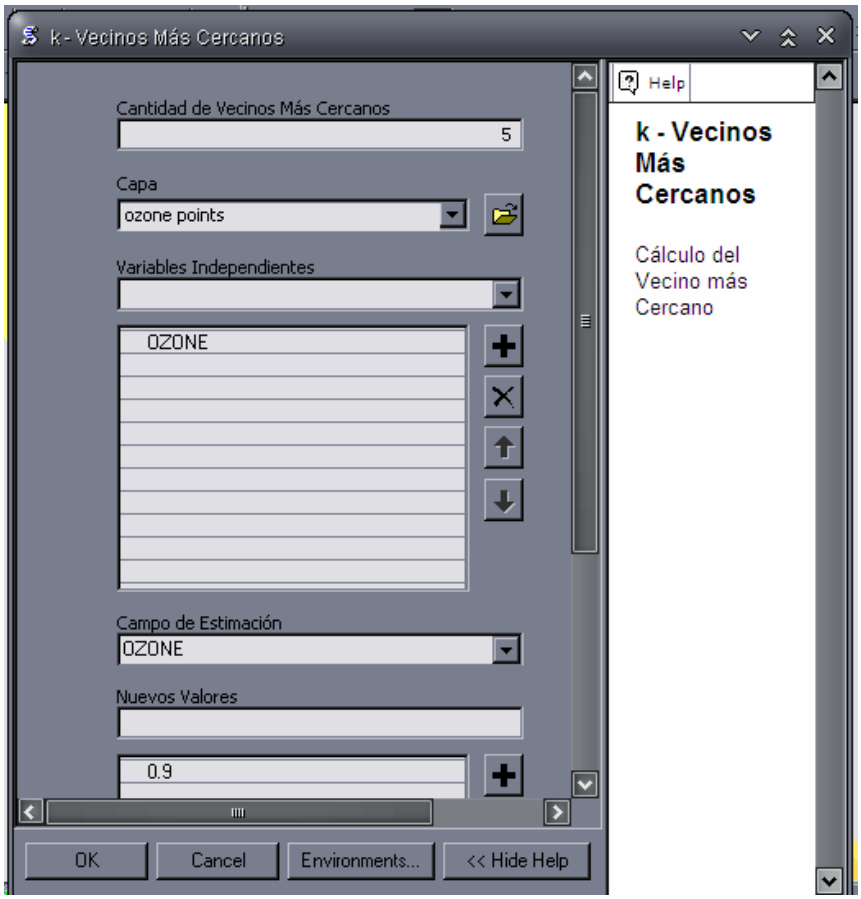


Figura 37 - Ventana de ingreso de cálculo de KNN

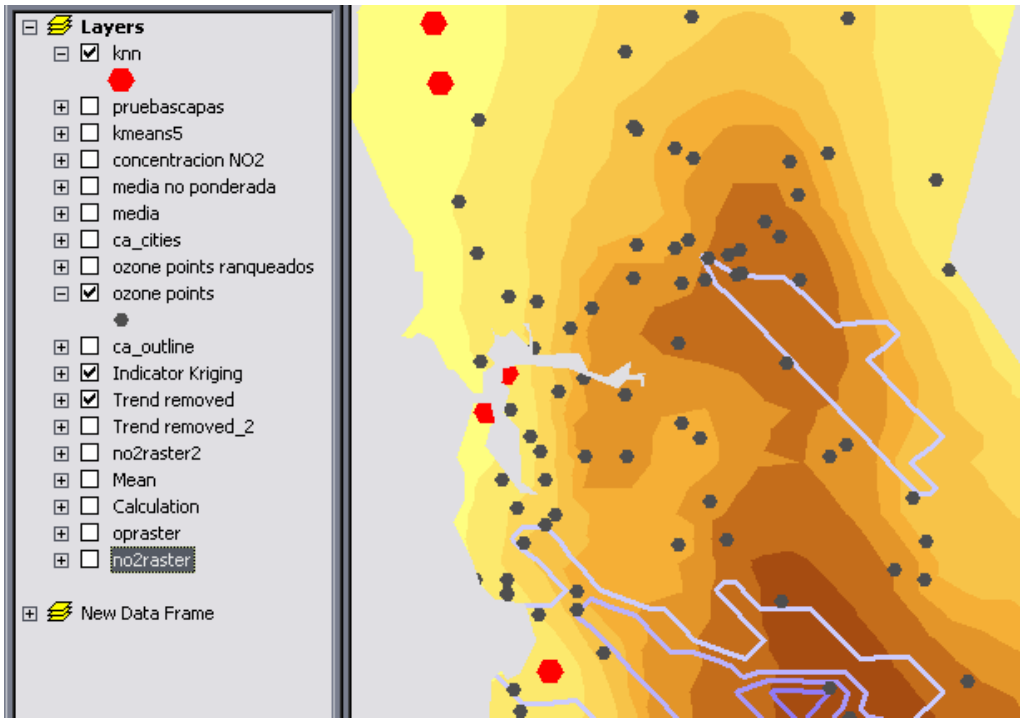


Figura 38 - Resultado del cálculo de KNN para 5 vecinos.

En la Figura 38 se muestra la capa resultante (en rojo) de los puntos "más cercanos" al valor ingresado de ozono. Como se ve, los puntos no tendrían porqué estar cercanos entre sí a no ser que existiera alguna condición geográfica que pudiera estar indicando algún problema en el área en cuestión.

Capítulo 5

Resultados Alcanzados

El producto obtenido es un prototipo de herramientas orientadas al estudio de epidemiología y minería de datos. El mismo es el resultado de una investigación exhaustiva de distintos métodos de enfoque para resolver el problema.

Entre las dificultades encontradas en el transcurso del proyecto la mayoría se debieron a periodos de tiempos muertos en donde se debían esperar respuestas y al problema que finalmente se tuvo que realizar el rol de epidemiólogo ya que no hubo contraparte a quien consultar a la hora de realizar las herramientas.

Finalmente se logró realizar un prototipo acorde a la necesidad de responder a la pregunta ¿Es posible realizar un producto que automáticamente controle todas las características de una epidemia genérica? Cuya respuesta es NO.

El prototipo de entrega final logró generar un *know how* suficiente para saber al menos que no existe una herramienta mágica para epidemiología ya que la diversidad de formas de solucionar un problema en epidemiología es tan grande como la cantidad de epidemiólogos hay para resolverla.

El conocimiento generado para herramientas futuras en el área y el enfoque a utilizar resultan en un gran aporte si se quisiera realizar un conjunto de herramientas comerciales para su uso en epidemiología.

Trabajo Futuro

Las extensiones al trabajo realizado son varias y de variada índole. Una posible extensión es el de generarse herramientas más específicas para estudios particulares y así automatizar tareas de contról.

Podría además incursionarse en el tema de simulación que fue inicialmente dejado para una posible extensión ya que por su envergadura superaba los tiempos y dimensiones del proyecto.

Criticas al Proyecto

Básicamente no existen grandes diferencias entre lo planeado y lo logrado. Pero existieron distintos contratiempos que resultaron en un atraso generalizado del proyecto y que no permitieron la aceptación de lo realizado por expertos en el área. Cosa que potencialmente podría llevar a que el estudio de las herramientas analizadas sea fútil.

Como primer punto en contra podemos ver la falta de un cliente experto en el área debido a las razones anteriormente expuestas. Esto influyó en que lo realizado no pudiera ser validado más allá que se haya buscado material extra que permitiese dar una idea de lo que resultaría interesante y útil tener como herramienta de análisis epidemiológico. Además, y como no se podía consultar a expertos cuando se tenían dudas sobre un cierto tipo de estudios o herramienta de análisis, se ocasionaron retrasos a la hora de analizar e implementar porque se necesitaba recurrir a material extra para poder buscar y estudiar todo lo posible sobre estos.

Otro punto es el que no se pudo lograr una herramienta "automática" de alerta epidemiológica porque no es posible generar un consenso de expertos en el área que dijeran como hacer el estudio tipo. De todas formas se pudieron crear simulaciones "a mano" (ejemplos controlados) que muestran como en casos particulares se podrían realizar herramientas de control automático de alerta temprana.

Como tareas a desarrollar primero debería ser el tener un grupo de expertos en el área de la Salud Pública que determine que y cuantas herramientas son útiles para generar un Sistema de Alerta Epidemiológica destinado a Uruguay así como que y cuantos datos y de donde obtenerlos y como reportarlos para su mejor análisis y uso para toma de decisiones. Yluego, basandose en el modelo realizado de herramientas, generar aquellas que los expertos consideren más convenientes.

Referencias

[1] http://www.fisicanet.com.ar/biografias/cientificos/p/pasteur.php, último ingreso 23/09/08

[2] http://www.ph.ucla.edu/epi/Snow/snowbio.html, último ingreso 23/09/08

[3] http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=1184, último ingreso 23/09/08

[4] http://ais.paho.org/sigepi/index.asp, último ingreso 23/09/08

[5] http://www.cica.es/epiinfo/, último ingreso 23/09/08

[6] http://www.ica.com.uy/empresa.asp?item=1, último ingreso 30/09/08

[7] http://www.esri.com/software/arcgis/index.html, último ingreso 15/10/08

[8] http://www.esri.com/, último ingreso 15/10/08

[9] http://www.elgeomensor.cl/downloads/manuales%20y%20tutoriales/index.php?file=Curso_ArcMap.doc, último ingreso 18/10/08

[10] http://www.python.org/, último ingreso 18/10/08

[11] McDuffie, Tina Spain (2003). *JavaScript Concepts & Techniques: Programming Interactive Web Sites*. Franklin, Beedle & Associates. ISBN 1-887-90269-4.

[12] http://msdn.microsoft.com/en-us/vbasic/default.aspx, último ingreso 18/10/08

[13] Documentos de FAO - Los sistemas de información geográfica y la tele percepción en la pesca, capitulo 6.2 - www.fao.org/DOCREP/003/T0446S/T0446S07.htm, último ingreso 24/09/08

[14] Modeling Our World: The ESRI Guide to Geodatabase Design; Michael Zieler; ESRI, Inc., 1999; ISBN: 1879102625, 9781879102620 - capítulo 3

[15] Epidemiology: Principles and Methods; Brian, M.D. Macmahon, Dimitrios Trichopoulos; Lippincott Williams & Wilkins, 1996; ISBN-13: 978-0316542227

[16] Discovering Data Mining: From Concept to Implementation; Peter Cabena; Prentice Hall, 1997; ISBN: 0137439806, 9780137439805

[17] http://www.w3.org/2002/ws/, último ingreso 18/10/08

[18] http://www.ganttchart.com/, último ingreso 18/10/08

[19] http://www.geotecnologias.co.cr/ESRI_9vr/ArcGIS_Desktop.asp, último ingreso 18/10/08

[20] http://java.sun.com/, último ingreso 18/10/08

[21] http://*www.esri.com/library/whitepapers/pdfs/shapefile.pdf*, último ingreso 18/10/08

[22] http://www.fisterra.com/mbe/investiga/var_cuantitativas/var_cuantitativas.asp, último ingreso 18/10/08

[23] http://www.wxpython.org/, último ingreso 18/10/08