Desarrollo de una herramienta para automatizar la estimación de datos faltantes en informes meteorológicos

Matías Antúnez Luciano Gervasoni Emmanuel Maggiori
Amalia Thomas

{mantunez,lgervasoni}@alumnos.exa.unicen.edu.ar
{emaggiori,athomas}@alumnos.exa.unicen.edu.ar

Dirección

Adriana Basualdo, Gustavo Illescas, Daniel Xodo

Universidad Nacional del Centro de la Provincia de Buenos Aires

Trabajo de Cátedra: Investigación Operativa

*****

Development of a tool that automates the estimation of missing data from climate observation stations

Abstract

In this paper we describe the development of a piece of software that automates the estimation of climate data. It was aimed at choosing automatically, in every estimation, the best interpolation method under the assumption that there is no "universal" technique and hence a sensible selection of the most suitable method is needed every time. In addition, the later application of the method itself should be automatic, suppresing all interaction with the user to make decisions. Inverse-distance weighting and Kriging methods were included. We show the steps taken to fully automate the tasks.

In the tests done, the errors computed were highly satisfactory, and it was shown that choosing the method in every estimation was a good approach. It was observed that some indicators show tendencies toward certain interpolation methods to achieve a lower error, and a close relation between the efficiency of the tool and the availability of information was found.

*****

Desarrollo de una herramienta para automatizar la estimación de datos faltantes en informes meteorológicos

Resumen

En este trabajo se describe el desarrollo de una herramienta de *software* que automatiza la estimación de datos faltantes de informes meteorológicos del Servicio Meteorológico Nacional. Se realizó como trabajo final, concluida su cursada, de una materia en Investigación de Operaciones correspondiente al tercer año de una carrera de Ingeniería de Sistemas.

El objetivo de la herramienta es realizar de manera automática la elección del mejor método de interpolación para cada dato faltante, bajo la suposición de que no existe un método "universal" y por lo tanto es necesario evaluar en cada caso cuál es el más adecuado. La posterior aplicación del método seleccionado también debía ser automática, sin necesidad de interacción con el usuario para tomar decisiones.

Se consideraron cinco métodos de uso muy común en la meteorología: IDW -lineal y cuadrático- y Kriging -simple, ordinario y universal-.

En este trabajo se describen las decisiones que se tomaron para lograr la automatización total de la herramienta, y su eficacia se verificó haciendo un análisis de errores para datos de cuatro años consecutivos.

En las pruebas realizadas, el método elegido se alterna, validando que es adecuado elegir cada vez el mejor, sin presuponer superioridad de uno sobre otro. Se vio que algunos indicadores se inclinan más hacia determinados métodos que otros y se encontró una relación entre la disponibilidad de información y el error en las estimaciones.

1. Introducción

Es frecuente la ausencia de algunos datos en los informes que listan las mediciones provenientes de estaciones meteorológicas. La ausencia de estos datos se puede deber, por ejemplo, a fallas en los intrumentos de medición o a problemas de comunicación.

Cuando se confeccionan reportes basados en las mediciones es conveniente estimar los datos faltantes de acuerdo a la información proveniente de las otras estaciones. Para eso se emplean métodos de interpolación, que buscan estimar un dato faltante a partir de los datos existentes.

Existen diversos métodos de interpolación. A veces determinado método es adecuado en ciertas condiciones, mientras que otro puede serlo en otras, debido a la disparidad en la correlación espacial de los datos. Por ejemplo, las tormentas se caracterizan por producir lluvias fuertes muy localizadas mientras que las *nimbostratus* producen lluvias más parejas y en extensiones amplias.

Esta herramienta se realizó impulsada por una meteoróloga que trabaja en el Ministerio de Agricultura, para facilitar el trabajo que se hace a diario, manualmente, en la Oficina de Riesgos Agropecuarios. Actualmente se la está empleando de manera satisfactoria en dicha oficina. Además, se la facilitó para su uso en el Servicio Meteorológico Nacional (SMN).

La herramienta debía entonces evaluar cada método para una circunstancia determinada, elegir la mejor opción y realizar la estimación con el método seleccionado. La aplicación del método también debía ser automática, sin interactividad con el usuario, a diferencia de la mayoría de las aplicaciones comerciales que requieren la participación de éste para tomar decisiones durante la interpolación (ver [12] para un ejemplo concreto).

De esta manera, se contaría con una herramienta que toma como entrada informes del SMN para producir como salida un informe cuyos datos faltantes fueron rellenados procurando realizar la mejor estimación posible en cada caso.

2. Métodos de interpolación

Las técnicas de interpolación empleadas en este trabajo consisten en la asignación de pesos a todas las mediciones existentes, para efectuar la estimación como una combinación lineal de los datos según los pesos asignados. Por lo tanto, si el vector $X = (x_1, ..., x_n)$ corresponde a las mediciones, se calcula un vector de pesos $\lambda(Y)$, en función de Y (la medición faltante). El valor faltante Y se estima de la siguiente forma:

$$Y = \sum_{i=1}^{n} \lambda_i X_i$$

Si a la vez se cumple que $\sum_{i=1}^{n} \lambda_i = 1$, entonces se dice que los pesos no están sesgados (*unbiased*) o se habla de una asignación justa.

2.1. Ponderación inversa a la distancia (IDW)

El método IDW, por su sigla en inglés (*Inverse-Distance Weighting*), asigna los pesos de manera decreciente a medida que las estaciones se alejan del dato que se debe estimar. Es decir que cuanto más lejos esté una estación del punto en cuestión, se considera que menos influencia debe ejercer sobre la estimación. Por eso se habla de una ponderación inversa a la distancia:

$$\lambda_i(Y) = \frac{T}{dist(X_i, Y)^p}$$

Se añade la posibilidad de elevar a un exponente p, de manera tal de poder elegir con qué rapidez se desea que decrezca el peso. T corresponde a la inversa de la suma de todos los pesos de las mediciones con respecto a la faltante, de manera tal de ajustar la fórmula para que la sumatoria sea igual a 1.

Además de emplearse el método IDW lineal ($p = 1$), es muy común utilizar IDW cuadrático ($p = 2$) para representar la correlación espacial de precipitaciones y otros indicadores [2], y se encontró documentación que avala la superioridad de la versión cuadrática en algunas circunstancias [8]. Por esta razón, se implementaron ambos métodos, con posibilidad de extenderse fácilmente a otros valores de p.

La desventaja principal del método es la arbitrariedad del valor de p.

2.2. Kriging

El método Kriging busca primero efectuar un análisis de la correlación espacial de los datos. Para esto se crea un variograma (o semivariograma) que expresa cuánto varían los datos de las mediciones en función de la distancia entre ellas. Con los datos obtenidos del variograma se resuelve un sistema de ecuaciones lineales para encontrar los pesos óptimos. Este método busca, entonces, obtener los pesos con un conocimiento de cómo se relacionan los datos según la distancia, lo que evita la arbitrariedad de IDW. Si bien se ha demostrado la superioridad teórica de los métodos Kriging por sobre los IDW, éstos requieren interacción para ajustar una función al variograma. Por esta razón, en la práctica, Kriging compite con IDW al querer automatizarlo [5].

3. Método Kriging de interpolación

El método Kriging se aplica siguiendo una serie de pasos:

1. Crear un variograma (o semivariograma) a partir de los datos.
2. Agrupar los puntos del variograma en clases de acuerdo a las distancias.
3. Aproximar con una curva al modelo experimental obtenido.
4. Crear el sistema de ecuaciones correspondiente, rellenándolo siempre con datos provenientes del modelo matemático y no del experimental.
5. Resolver el sistema de ecuaciones para obtener los pesos y así interpolar.

A continuación se detalla cada uno de los pasos y se explican las decisiones de implementación y automatización que se llevaron a cabo en el desarrollo de la herramienta.

3.1. Creación del semivariograma empírico

El semivariograma se define como

$$\gamma(h) = \frac{1}{2 \mid N(h) \mid} \sum_{(i,j) \in N(h)} (X_i - X_j)^2$$

donde $N(h)$ es el conjunto de pares de estaciones i, j a distancia h. Se ve intuitivamente que la función representa el promedio del cuadrado de la variación entre estaciones a una distancia determinada. En otras palabras, se buscan todas las estaciones situadas a distancia h y se promedian sus varianzas.

Lo natural es que para valores pequeños de distancia h, la semivarianza $\gamma(h)$ sea baja también, ya que los datos serán más parecidos cuanto más cerca se encuentren. A medida que estos se van alejando, es también natural que la semivarianza vaya aumentando. En algún punto los datos dejan de estar relacionados entre sí, la curva se vuelve aproximadamente horizontal y la varianza se asemeja a la varianza de la muestra, ya que al no influenciarse las mediciones entre sí, ésta no depende de la distancia. A esta distancia se la conoce como **rango** del variograma, y a la semivarianza correspondiente como **silo** [4].

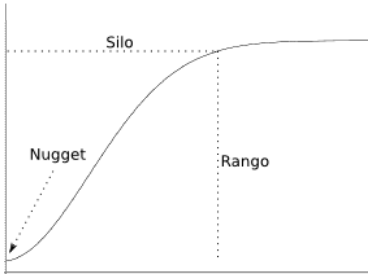

Figura 1: Modelo de variograma gaussiano

En la figura 1, se puede apreciar además el efecto ***nugget***, por el cual a distancia cero no es cero la varianza. Esto parece no ser natural en un principio, ya que no debería haber variación entre datos a distancia nula. Este efecto sucede por la posible discontinuidad de los datos a pequeña escala, por ejemplo, si se mide la presencia de un mineral en una mina, y hay dos piedras de distinta naturaleza contiguas.

En la práctica, para construir el semivariograma simplemente se itera sobre cada par de puntos (realizando entonces $O(n^2)$ operaciones) y se computa el valor correspondiente para almacenarlo en una estructura de datos.

3.2. Agrupamiento en clases

Los datos se suelen agrupar en clases. Por ejemplo, se crea una clase para las distancias de 1000 a 1200km y se representa la varianza como el promedio de las varianzas de todos los puntos que caen en esa clase. También se puede

hacer directamente en la construcción del variograma, aplicando una tolerancia a la distancia h.

El objetivo del agrupamiento en clases es disminuir la cantidad de información por una cuestión de tiempos de cómputo, necesario en este trabajo debido al alto costo temporal del proceso. Es discutible que a fines numéricos se mejore la precisión agrupando los datos [11].

El tamaño de cada clase debe ser elegido apropiadamente. Para automatizar esta elección, se siguió la recomendación de la documentación disponible en Internet de ayuda de la herramienta comercial *ArcGis*, donde se sugiere que se elija como tamaño de clase (*lag size*) al promedio de distancia al vecino más cercano. Para este fin, se aprovecha la contrucción del variograma (tarea que requiere realizar n^2 operaciones entre todos los pares de mediciones) y se registra de cada estación su más cercana, y luego se calcula el promedio entre los datos obtenidos. Al agrupar fue necesario conservar la información de cuántos puntos había originalmente en cada clase, ya que es necesario para un proceso posterior de regresión no lineal.

3.3. Aproximación de un modelo matemático

El modelo experimental tiene la desventaja de consistir en una nube de puntos, poco pareja, discontinua, con *outliers*. Además, se necesitan posteriormente datos que no están presentes en el modelo experimental. Por lo tanto, se debe aproximar una curva que se ajuste al modelo. Con frecuencia los modelos empleados para este fin son el esférico, exponencial y gaussiano, por lo que se incluyeron esas tres alternativas en este trabajo.

La herramienta desarrollada ajusta los tres modelos a la nube de puntos, mediante una regresión no lineal. A continuación elige el modelo que menor error aporta y emplea a ese modelo en el paso siguiente. De esta manera se elige la curva que mejor se adapta al variograma experimental.

Para realizar la aproximación de la curva se debía emplear algún algoritmo de regresión no lineal, por la cual iterativamente se busca el juego de parámetros de la función que disminuye el error entre la nube de puntos y la curva trazada. El algoritmo elegido fue *Levenberg-Marquardt*. Consiste en una combinación sensata de los algoritmos de *Gauss-Newton* y del gradiente. Inicialmente usa la técnica del descenso por el gradiente más pronunciado, para acercarse a la solución. Cuando está lo suficientemente cerca de ésta, el algoritmo se convierte en Gauss-Newton, para buscar con mayor precisión los valores óptimos [10].

Estos algoritmos son capaces de encontrar el juego de variables que produce un mínimo *local* para el estimador de error. Se corre el riesgo de que ese mínimo local no sea o esté muy alejado del mínimo *global* que se está buscando para obtener una buena aproximación de la curva. Por esa razón, es necesario que los datos iniciales de las variables (requeridos para comenzar a iterar por cualquiera de los algoritmos mencionados anteriormente) estén cerca de los valores óptimos.

Las fórmulas de los modelos implementados cuentan con tres constantes que deben ser obtenidas mediante la regresión. Esas constantes tienen una correspondencia directa con los elementos del semivariograma.

La primer constante de las ecuaciones corresponde al *nugget* y se decidió atribuirle un valor inicial igual a cero para iniciar las iteraciones, debido a que se sabe que su valor es cercano al nulo.

La segunda constante corresponde al silo parcial, es decir a la diferencia entre el silo y el *nugget*. Como estimamos un *nugget* inicial igual a cero, en este caso consideramos equivalentes al silo y al silo parcial. Se considera como valor inicial a la semivarianza de la muestra ($\sigma^2/2$) dado que se supone que cuando los datos no se influencian los unos a los otros (cuando la distancia es mayor al rango), el valor que toma la varianza es aproximadamente similar al valor de la varianza de la muestra [1].

La última constante corresponde al rango. Se vieron sugerencias de estimar, por ejemplo, que éste se encuentra a la mitad del variograma según el eje horizontal. Ese tipo de estrategias, según se vio, hacían que la regresión cayera en mínimos locales. Se decidió, por lo tanto, emplear un criterio haciendo uso del llamado rango efectivo, que es el primer valor de distancia h que alcanza el 95 % del silo [3]. Para ésto, se recorre el variograma de izquierda a derecha y cuando se encuentra el primer punto que cumple con la condición, se establece a esa abscisa como el valor inicial para el rango en la iteración.

El algoritmo de regresión no lineal busca encontrar un mínimo para un estimador de error. Normalmente se emplea como indicador al error cuadrático medio. Sin embargo, para aproximar variogramas este método no es correcto ya que ignora las covarianzas entre los datos. Se puede emplear un estimador que aproxima mejor el error (criterio de Cressie) [6], definido como

$$E(\theta) = \sum_{j=1}^{k} \mid N(h_j) \mid \left(\frac{\hat{\gamma}(h_j)}{\gamma(h_j, \theta)} - 1 \right)^2 ,$$

donde $\mid N(h_j) \mid$ es la cantidad de puntos contenidos en la clase (por lo que se debía conservar esa información al agrupar), $\hat{\gamma}$ es el semivariograma experimental, γ el modelo matemático y θ un vector con los parámetros de la función en cuestión.

Se observó en las pruebas que usando este criterio se mejoraba notablemente la aproximación realizada.

3.4. Sistemas de ecuaciones

Las tres variantes de Kriging que se implementaron difieren en este paso.

El método Kriging Simple es el más sencillo de los tres y supone que el espacio es estacionario (donde la media no varía en función de la posición) y no garantiza la imparcialidad (*unbiasedness*) por lo que los pesos asignados podrían no sumar uno. Se debe resolver el siguiente sistema de ecuaciones:

$$\begin{pmatrix} \gamma(h_{1,1}) & \gamma(h_{1,2}) & ... & \gamma(h_{1,n}) \\ \gamma(h_{2,1}) & \gamma(h_{2,2}) & ... & \gamma(h_{2,n}) \\ ... & ... & ... & ... \\ \gamma(h_{n,1}) & \gamma(h_{n,2}) & ... & \gamma(h_{n,n}) \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \gamma(h_{1,p}) \\ \gamma(h_{2,p}) \\ ... \\ \gamma(h_{n,p}) \end{pmatrix}$$

La función γ empleada es el modelo matemático y no el semivariograma experimental. La matriz está formada por la evaluación en esa función de las distancias entre todos los pares de mediciones. El vector columna se contruye a partir de las distancias entre el punto p que se desea interpolar, y todos los demás puntos.

Se acostumbra a incluir en el sistema de ecuaciones (y en la posterior interpolación) sólo los puntos cuya distancia al dato faltante es inferior al rango, ya que según el modelo si los puntos están más alejados que el rango no tienen influencia.

El Kriging Ordinario incluye la restricción de justicia (que los pesos sumen uno), problema que se resuelve añadiendo un multiplicador de Lagrange.

El método Kriging Universal no asume que el espacio es estacionario. Esto suele ser el caso en los indicadores meteorológicos: la media de los datos tiene tendencias, por ejemplo, las precipitaciones se reducen a medida que se aleja de la costa, por lo que la media no es estacionaria. En este trabajo se consideró una tendencia lineal en los datos, y se agregan tres restricciones al sistema de ecuaciones para este fin [7].

El sistema de ecuaciones se resuelve mediante el Método de Eliminación de Gauss [13], siendo $O(n^3)$ en tiempo.

4. Validación cruzada

Como se deseaba que el sistema interpole automáticamente cada indicador en cada fecha de los archivos de entrada, se debía elegir qué método ofrece mejores resultados para cada uno de ellos (entre IDW Lineal, IDW Cuadrático, Kriging Simple, Kriging Ordinario y Kriging Universal).

Se eligió la validación cruzada (*cross-validation*) [9] como técnica para ponderar la calidad de cada método de interpolación. En particular, la "validación cruzada dejando uno fuera" (*Leave-one-out cross-validation* - LOOCV). Por turnos se deja afuera a una de las estaciones (la estación de evaluación) y se estima su dato mediante las $n - 1$ estaciones restantes (de entrenamiento). Los datos estimados se comparan con los datos reales y se acumula el error obtenido. Es decir, se supone faltantes a datos que no lo son, para evaluar la precisión del método en ese contexto. Esto se realiza con cada uno de las n estaciones. De esta manera se cuenta con dos significativas ventajas: se emplean todas las mediciones y a la vez todas toman tanto el rol de entrenamiento como el rol de evaluación.

La validación cruzada se realiza con cada método de interpolación implementado (para cada indicador en cada fecha), y se lleva finalmente a cabo la interpolación de los datos faltantes, según el método que haya aportado el menor error tras este procedimiento, lo que constituye un costo temporal $O(n^4)$. Se creó un hilo (*thread*) para cada método que se desea ponderar, aprovechando las ventajas del paralelismo en las máquinas que lo permiten. Los tiempos de ejecución resultaron totalmente satisfactorios según los especialistas.

5. Resultados

La herramienta resultante interpola satisfactoriamente los datos faltantes de las mediciones. Los métodos de interpolación elegidos por el algoritmo se van alternando: para el mismo indicador en distintas fechas se eligen métodos distintos, así como también el método varía para cada indicador de una fecha determinada.

Se realizaron pruebas con datos de cuatro años consecutivos (2009–2012), lo que constituye alrededor de 145.000 mediciones para cada indicador.

En la tabla 1 se muestran los errores medios que se registraron en el procedimiento de validación cruzada de las pruebas realizadas, para cada indicador, a lo largo de todo el período. Además, se computaron los errores desglosados por cada año, por cada estación.

	T. Máx.	**T. Mín**	**Heliof.**	**Vien. Máx.**	**Pp.=0**	**Pp.≠0**
EMA	1,5	1,9	1,5	11,8	0,6	4,4
EMA*	6,3 %	8,7 %	13,1 %	14,7 %	1,3 %	23,4 %
Dist.	208,8	220,6	213,3	236,6	207,3	
Cant.	92,6	93,7	77,7	75,6	100,9	

Tabla 1: Errores para cada indicador, donde EMA es el error medio absoluto y EMA* corresponde al mismo dato luego de una normalización.

En la primera fila se pueden ver los errores medios absolutos según la unidad de cada indicador. Por ejemplo, para la temperatura máxima se registró un error medio de 1,5 grados celsius. En el caso de la precipitación, se separaron los cómputos para los casos en que la precipitación era cero y distinta de cero (ya que hay una gran cantidad de datos nulos y se acostumbra esta separación). Las magnitudes cumplen con las expectativas del desarrollo de esta herramienta, según determinaron los especialistas. En la segunda fila, para poder comparar las calidades de la interpolación entre los disintos indicadores, se normalizaron los datos, dividiendo por el tamaño del intervalo de valores medidos en cada caso. Se puede interpretar, por ejemplo, que el error que se encontró para la temperatura máxima fue en promedio un 6,3 % del espectro de valores que tomó el indicador en cada fecha.

Es conveniente destacar que los datos son dispares en cuanto a la cantidad de información disponible para cada indicador, además de variar ésta en el tiempo. Por eso es de utilidad observar los datos desglosados, o bien tomar en cuenta la disponibilidad de información, que se ponderó mediante la distancia promedio al vecino más cercano y la cantidad promedio de mediciones (como se ve en las siguientes dos filas de la tabla).

Como era de esperarse, la lluvia presentó el mayor error porcentual. Para la heliofanía y el viento se hicieron estimaciones menos precisas que en los casos de la temperatura. Se sospechó que eso se debía a que para estos dos indicadores se cuenta con menos disponibilidad de información. Para verificarlo se evaluó la relación entre el error y la distancia promedio al vecino más cercano, y entre

el error y la cantidad de datos disponibles. Se corroboró que efectivamente a mayor distancia entre estaciones, mayor error, y a mayor cantidad de datos, menor error (como se presenta de manera sintética en la figura 2).

En el análisis desglosado por estación se apreció que el error para las precipitaciones es mayor en verano y primavera que en otoño e invierno, lo cual era lo esperable debido a la naturaleza convectiva de las lluvias en el primer caso.

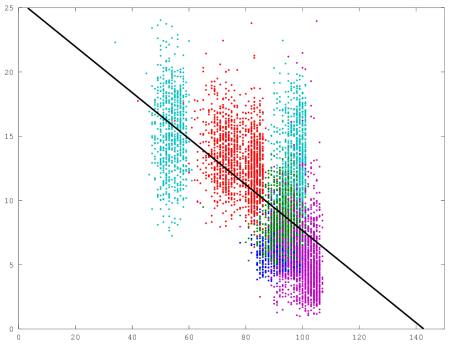

Figura 2: Error porcentual en función de la cantidad de mediciones disponibles

En la tabla 2 se lista el porcentaje de veces que se eligió cada método para cada indicador, según las pruebas realizadas. Para medir el grado de inclinación

	T. Máx.	T. Mín.	Heliof.	Viento	Pp.	**Promedio**
IDW Lineal	0 %	0,2 %	1,4 %	38,2 %	7,4 %	**9,44 %**
IDW Cuad.	25,2 %	47,6 %	22 %	46,6 %	39,4 %	**36,16 %**
Kr. Simple	1,6 %	1,5 %	18,3 %	3,7 %	26,1 %	**10,24 %**
Kr. Ord.	40,2 %	24,8 %	43,9 %	9 %	22,8 %	**28,14 %**
Kr. Univ.	33 %	25,9 %	14,4 %	2,5 %	4,3 %	**16,02 %**
Desvío	**18**	**20**	**15**	**21**	**14**	

Tabla 2: Resultados de la elección de métodos para los archivos de prueba

hacia determinado indicador, se incluyó el dato del desvío estándar para los porcentajes de cada método: a mayor desvío, menos justa la competencia entre métodos para el indicador.

Se vio que para los casos de la temperatura existe una inclinación hacia los métodos Kriging Ordinario y Universal, y hacia IDW Cuadrático. Es notable que IDW Lineal no se eligió ni una vez en los cuatro años para temperatura máxima. En el caso del viento, se ve una marcada tendencia a los métodos IDW, tanto cuadrático como lineal, donde este último compite muy justamente. En los casos de precipitación y heliofanía se da la elección menos dispar entre los métodos, en especial en las precipitaciones. En conjunto, los métodos Kriging representan más de un 50 % de las elecciones.

También se hicieron pruebas para comparar los errores obtenidos con el error producto de haber aplicado el mismo método siempre, sin elegir, durante los cuatro años. El método seleccionado para cada caso fue el que mejor éxito tuvo en las pruebas (según la figura 2). Se vio que el error aumenta en algunos casos muy notoriamente. Por ejemplo, la temperatura máxima pasa a tener un error promedio absoluto de 6,6 grados celsius en los cuatro años (contra 1,5 si se elige el mejor método cada vez). En algunos casos es aun más llamativo: por ejemplo, si se realiza este procedimiento sobre el verano de 2009, se obtiene un error de 11,2 grados contra 1,8 para ese mismo período.

6. Conclusiones

Lo más desafiante del desarrollo de la herramienta fue la eliminación de la interactividad con el usuario, ya que muchas tareas que se suelen realizar "a ojo" se debían automatizar. Para esto se emplearon técnicas como la validación cruzada para ponderar la calidad de los métodos de interpolación o un algoritmo iterativo de regresión no lineal para aproximar un modelo experimental con una curva matemática. También se debieron elegir maneras sensatas de estimar los valores iniciales de la regresión y para determinar el tamaño de clase de los variogramas, parámetros que suelen solicitarse interactivamente.

La herramienta resultante interpola satisfactoriamente los datos faltantes en archivos provenientes del SMN, según se observó a partir de contabilizar los errores obtenidos. Los errores, así como el tiempo de ejecución (mejorado mediante el uso de paralelismo) son totalmente admisibles.

Se ve que algunos indicadores se inclinan marcadamente hacia algunos métodos mientras que otros muestran menos tendencia hacia un método en particular. En el primer grupo entran las temperaturas y el viento máximo, y en el segundo las precipitaciones y la heliofanía. Aun el método que menos se elige en general, IDW Lineal, representa una de las elecciones favoritas en el caso del viento máximo, por lo que se justifica tomarlo en cuenta.

Se hicieron pruebas para comparar los errores obtenidos en el análisis, con el error producto de haber aplicado el mismo método siempre, sin elegir, durante los cuatro años. Se vio que el enfoque de este trabajo mejora los errores, a veces, de manera muy significativa. También se apreció que el error y la cantidad de datos disponibles (así como la distancia entre las estaciones) están relacionados.

Estos resultados validan el objetivo de este trabajo, que era construir una herramienta que elija el método más apropiado para interpolar cada dato, sin presuponer superioridad de un método sobre otro.

Referencias

[1] Randal J Barnes. The variogram sill and the sample variance. *Mathematical Geology*, 23(4):673–678, 1991.

[2] Adriana B. Basualdo. *Obtención y análisis de campos de correlación espacial de la precipitación en dos regiones de latitudes medias*. Tesis de grado, Departamento de Ciencias de la Atmósfera, UBA, Agosto 1999.

[3] Ling Bian. Modeling the sample variogram. Material de clase de Geo479/579: Geostatistics, University at Buffalo, State University of New York. Dispolible online. `www.geog.buffalo.edu/~lbian/ch16.ppt`.

[4] Geoff Bohling. Introduction to geostatistics and variogram analysis. *Kansas geological survey, 20p*, 2005.

[5] DS Bromberg, F Pérez. Interpolación espacial mediante aprendizaje de máquinas en viñedos de la Provincia de Mendoza, Argentina. In *13th Argentine Symposium on Artificial Intelligence, ASAI*. 41 JAIIO, 2012.

[6] Robert M Brunell. An automatic procedure for fitting variograms by cressie's approximate weighted least-squared criterion. *Department of Statistical Science Technical Report No. SMU/DS/TR, Southern Methodist University*, 1992.

[7] Nicolas Christou. Universal kriging. Material de clase de Statistics C173/C273, University of California, Los Angeles. Dispolible online, November 2011. `http://www.stat.ucla.edu/~nchristo/`.

[8] Louis de Mesnard. On inverse distance weighting in pollution models. Dispolible online, November 2011. `http://ssrn.com/abstract=1931636`.

[9] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[10] Manolis IA Lourakis. A brief description of the levenberg-marquardt algorithm implemented by levmar. *Institute of Computer Science, Foundation for Research and Technology*, 11, 2005.

[11] Werner G Müller. Least-squares fitting from the variogram cloud. *Statistics & probability letters*, 43(1):93–98, 1999.

[12] Ayse Irmak Parikshit Ranade and David R. Maidment. Geostatistical analyst, space interpolation methods. Dispolible online. `www.ce.utexas.edu/prof/maidment/giswr2008/geostat/ExGeostat.doc`.

[13] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*, chapter 2.2. Cambridge University Press, 2007.