pp. 205-214 (1994)

RIESGOS DE INTERPRETACIÓN DE LA MATRIZ DE CORRELACIÓN LINEAL EN CLIMATOLOGÍA. VARIABLES DE FACTOR PURO E IMPURO

José Manuel SÁNCHEZ MARTÍN
Investigador en el Dptº de Geografía y O. T.
Universidad de Extremadura. Cáceres.

RESUMEN: La matriz de correlación lineal es una de las técnicas más utilizadas en Climatología. Sin embargo, pensamos que este método se ha aplicado de forma errónea. Por lo tanto, en este artículo hacemos unas breves reflexiones acerca de los riesgos que presenta esta técnica e intentamos dar una solución a los problemas propuestos.

ABSTRACT: The linear correlation matrix is one of the most widely used statistical technique in Climatology. However, we think that this methods is applied and interpreted badly. Therefore, in this article some brief reflections about the risks of this technique are presented and a solution to the problems involved is proposed.

La matriz de correlación es una de las técnicas estadísticas utilizadas con más profusión en la Climatología actual, debido a los buenos resultados aparentes que ofrecía, sobre todo cuando se correlacionaban variables geográficas con las climáticas.

Mediante esta correlación se obtiene la forma en que covarían dos variables, sin intentar profundizar más allá. Debido a ello, nosotros proponemos que no se establezcan sólo diferencias entre coeficientes de correlación causales y casuales.

Los primeros, obedecen a la presencia de una relación causa-efecto, en la que cualquier cambio en una variable está íntimamente ligado al que se produce en otra. Este tipo de correlación no siempre es visible, por lo que sólo puede ser el investigador el que decida si realmente las dos variables relacionadas guardan una dependencia directa. Con ello queremos decir que los coeficientes de correlación deben ser supervisados por el porpio investigador. Será él quien, en definitiva, decida si la covariación está provocada por un fenómeno causal.

Los segundos están motivados por simples coincidencias. De ese modo es

posible la obtención de unos coeficientes muy elevados, pero que no son debidos a un relación causal. Por lo tanto, estos valores deben ser ignorados por el investigador a la hora de interpretar los resultados.

Estas reflexiones preliminares no son nuevas, se han podido ver en la literatura existente al respecto. Sin embargo, era necesario incidir en ellas, sobre todo teniendo en cuenta que se trata de un análisis de una técnica estadística compleja, la matriz de correlación simple o lineal.

Por el contrario, cuando nos referimos al coeficiente de correlación mínimo para considerar una relación lineal como significativa, nos encontramos con la disparidad manifiesta que existe. Este valor fluctúa entre el 0.300 y el 0.350, lo que implica un porcentaje de explicación de la varianza cifrado entre un 9% y un 12.25%.

Si aceptamos estos valores es necesario tener en cuenta que el grado de solapamiento entre los diferentes coeficientes de correlación no es muy significativo, pues de un teórico 100% que se obtendría con una correlación pura, tan sólo rondamos el 10%.

Debido a este grave inconveniente, no es posible establecer afirmaciones más rotundas, fiables y taxativas. Ello se traduce, sobre todo en los no iniciados en el análisis estadístico, en una duda con respecto a esta técnica.

Sin embargo, si incrementamos el coeficiente de correlación de significación mínima hasta situarlo en 0.500, estableceremos un umbral crítico de explicación de la varianza de una variable con respecto a la otra que se cifra en el 25%. Este valor ya es mucho más adecuado para permitirnos establecer una serie de afirmaciones, con un grado de fiabilidad muy superior al anterior.

Pese a ello, debemos reconocer que, en muchas ocasiones, el coeficiente de correlación que se obtiene para una muestra grande, suele ser inferior que cuando se trata de una serie pequeña. Entonces, siguiendo a determinados autores (Raso, Martín, Clavero) es precisorecurrir a una formulación matemática que nos permita conocer el coeficiente de correlación crítico para conceder un cierto grado de importancia. Su fórmula consiste en:

$$r\sqrt{n} > 1.96$$

No obstante, con esta razón matemática no se obtienen resultados demasiado brillantes, sobre todo si tenemos en cuenta que con una muestra de 100 individuos, tan sólo sería necesario un coeficiente de correlación superior a 0.196. Este valor se traduce en un 3.84% de explicación de la varianza. Se trata, indudablemente, de un porcentaje de explicación muy pobre. Debido a este

motivo, no creemos que sea la opción más adecuada, sobre todo cuando se analiza la recta de regresión a que da lugar este coeficiente. Si la realizamos, nos percataremos de la gran dispersión que existe entre los diferentes puntos muestrales y la recta construída al efecto.

Por todo ello, pensamos que puede ser una solución no del todo adecuada, máxime si tenemos en cuenta el elevadísimo porcentaje de la varianza que queda sin explicar.

En cambio, nosotros proponemos que se utilicen unos coeficientes de correlación más elevados, para poder realizar un análisis más riguroso. Por lo tanto, creemos que son acertados los siguientes taxones, obivando el signo:

Explicación baja: entre un 15 y un 25%	r= 0.387 < 0.500
Explicación moderada: entre un 25 y un 50%	r= 0.500 > 0.707
Explicación buena: entre un 50 y un 75%	r= 0.707 > 0.866
Explicación muy buena: superior a un 75%	r= > 0.866

Con estos coeficientes de correlación es posible realizar un análisis bastante preciso de la relación que se establece entre las diferentes variables que intervienen en el cálculo estadístico.

Teniendo en cuenta la tipificación que proponemos, podrá observarse que nos mostramos partidarios de la elección de unos valores muy elevados. Este hecho es comprensible por que mientras más importante sea la covariación de dos variables, tanto más lo será la explicación que podamos efectuar sobre el particular.

De este modo, pensamos que con un coeficiente de correlación inferior a 0.500, se consigue una explicación de la varianza baja. Esta afirmación es comprensible, ya que se queda sin explicar un 75% de la misma.

Por el contrario, cuando utilizamos coeficientes superiores a 0.500, pero inferiores a 0.707, explicamos un porcentaje que se cifra entre el 25 y el 50%. Como vemos, no es una cifra extremadamente elevada, ya que nuevamente tendremos un elevado porcentaje de la varianza intrínseca de las dos variables sin explicar.

En cambio, si los valores de correlación son incrementados hasta dar una explicación de la varianza que se sitúe entre el 50 y el 75%, tendremos que recurrir a coeficientes tan elevados como los que van desde 0.707 a 0.866. Con estos valores es posible efectuar unas afirmaciones mucho más fundamentadas, con menor riesgo de error, pese a que la dispersión entre los puntos de observación y la recta de regresión continúa siendo elevada.

Por último, proponemos que se tomen como correlaciones muy buenas

aquéllas que aporten una explicación de la varianza superior al 75%, es decir, que el coeficiente de correlación sea superior a 0.866. Si tomamos este valor, tendremos una posibilidad de acierto mucho mayor que la obtenida con una menor correlación, siendo posible efectuar afirmaciones más rotundas.

Como puede observarse, tomamos como valores adecuados unos coeficientes de correlación lineal muy significativos, por lo que es posible pensar que en pocas ocasiones pueden obtenerse. No obstante, no es así, ya que uno de los riesgos primordiales de la matriz de correlación lineal consiste en conformarse con la consideración de coeficientes significativos unos valores muy bajos. Con ello, pensamos que se comete un error, ya que si se adopta una covariación excesivamente baja, estaremos limitados en nuestras deducciones y, además, despreciaremos un porcentaje de irregularidad que es muy significativo.

Es éste, pues, uno de los errores más comunes que se comete al analizar una matriz lineal. De igual manera, puede contribuir a un equívoco aún mayor si aplicamos la misma forma de proceder a otras técnicas estadísticas complejas, entre las que se cuenta cualquier tipo de análisis factorial, que en sus fundamentos básicos utiliza la matriz de correlación.

Teniendo en cuenta estas afirmaciones, llegamos a la conclusión de que resulta totalmente necesario mostrarse riguroso a la hora de establecer los coeficientes de correlación significativos, para no cometer ciertas irregularidades en la interpretación, al asimilar una covarianza reducida a dos variables.

Pese a que todos estos aspectos son de especial interés para llevar a cabo una buena interpretación de los resultados que arroja la matriz de correlación, no debemos menospreciar otro, como es la dificultad de correlacionar variables de factor puro con otras de factor impuro. Dicho de otra forma, no es fácil elaborar una matriz con variables simples y complejas, ya que los fundamentos interpretativos diferirán notablemente, al igual que los resultados.

Debido a esto, es preciso aclarar o definir las variables de factor puro y las de factor impuro, esto es, las simples o las complejas, respectivamente.

De este modo, es posible asimilar las variables de factor puro a aquéllas que no están modificadas por otras. Serían las variables independientes, que influyen en otras variables pero ellas, en sí mismas, no son modificadas por ninguna.

En cambio, las variables complejas o de factor impuro, resultan mucho más difíciles de analizar, debido a que en su configuración o en su alteración intervienen diferentes variables de factor puro. En definitiva, se trataría variables dependientes, pero con una característica fundamental, estarían alteradas no sólo por un factor independiente, sino por varias. Teniendo en cuenta estas apreciaciones, nos percataremos de que no siempre es fácil

conseguir relacionar una variable simple con otra compleja. Se debe a que la covariación no puede establecerse de forma idónea entre ellas, puesto que una técnica bivariante jamás podrá realizar un análisis multivariante óptimo de éstas. Para ello se precisa una técnica multivariante.

Con el fin de ejemplificar estas aseveraciones vamos a recurrir a una variable climática, las precipitaciones. Mediante ellas será posible observar el problema que surge cuando aplicamos una matriz de correlación lineal con los factores geográficos.

Así mismo, es preciso señalar que partimos de un conjunto de 72 observatorios climáticos de la red estándar que cubre Extremadura y, para aplicar la matriz de correlación hemos tomado cinco factores geográficos, los principales. Se trata de la altura, el emplazamiento (altura relativa), la latitud, la longitud y, por último, la exposición.

Con la inclusión de todos los datos, obtenemos una matriz de correlación, elaborada siguiendo las pautas "normales", es decir, la interpretamos de la forma tradicional.

CUADRO 1

VAR. CLIMÁT.	ALTURA	EMPLAZ	LATITUD	LONGIT.	EXPOSI
P. ENERO	0.310	0.600	0.660	-0.060	0.550
P. FEBRERO	0.290	0.670	0.710	-0.200	0.500
P. MARZO	0.220	0.600	0.740	-0.110	0.400
P. ABRIL	0.450	0.590	0.580	-0.030	0.490
P. MAYO	0.270	0.750	0.810	-0.220	0.330
P. JUNIO	0.280	0.710	0.660	-0.160	0.290
P. JULIO	0.110	0.510	0.470	-0.100	0.130
P. AGOSTO	0.170	0.530	0.600	-0.150	0.140
P. SEPTIEMB.	0.210	0.630	0.740	-0.170	0.310
P. OCTUBRE	0.330	0.650	0.690	-0.090	0.510
P. NOVIEMBRE	0.360	0.560	0.670	0.050	0.480
P. DICIEMBRE	0.320	0.680	0.640	-0.100	0.520
P. ANUAL	0.330	0.680	0.730	-0.100	0.540

Si analizamos detenidamente esta matriz de correlación, observamos que los coeficientes obtenidos entre las precipitaciones y la altitud no son muy elevados,

apenas se supera el 10% en la explicación de la varianza, si bien la situación varía en función de los diferentes meses del año.

Este hecho no deja de sorprendernos, sobre todo cuando buena parte de los postulados físicos demuestran un aumento pluviométrico considerable con la altura. Sin embargo, es posible dar una explicación lógica como es la presencia de observatorios en zonas bajas, en valles de montaña. Este hecho contribuye a la dispersión de los datos con respecto a la recta de regresión definida por dicho coeficiente.

En cambio, el emplazamiento o la altura relativa de un observatorio, guarda una relación mucho más importante con las precipitaciones que la altura absoluta. De este modo hemos calculado una covariación media de casi un 30%, cifra mucho más que importante si seguimos a muchos autores que abogan por el umbral crítico de 0.350. El coeficiente de correlación tan elevado es explicable por el efecto de pantalla orográfica que ofrece un sistema montañoso a las áreas cercanas, si bien esto no es lo que más nos interesa, sino el coeficiente obtenido.

Por su parte, el grado de correlación calculado para las precipitaciones y la latitud es también muy elevado, ya que durante algunos meses supera el 65% de explicación de la varianza, siendo la media inferior al 45%. Esta cifra es comprensible si tenemos en cuenta que los frentes que afectan Extremadura tienen una componente oeste. Por lo tanto, siguen un desplazamiento específico que convierte a la zona septentrional en mucho más húmeda. A esto hay que añadir otro aspecto geográfico importante, los mayores relieves se encuentran en el norte, lo que naturalmente redunda en un incremento pluviométrico.

La longitud, por el contrario, mantiene un escaso grado de relación con las precipitaciones de la mayor parte de los meses del año. Además, se trata de un coeficiente de carácter negativo, lo que implica una cierta anomalía. Es decir, de estos datos se deduce que a mayor longitud, más al oeste por tanto, se produce una cantidad de precipitación menor. Este hecho resulta, *per se*, poco creíble en una zona como Extremadura, ya que no tendría validez la influencia oceánica que posee la zona occidental de la misma. Teniendo en cuenta que este hecho es poco aceptable en esta área, es preciso revisar los datos de partida.

Cuando lo hacemos, nos percatamos de la especial distribución de los observatorios que forman la muestra. Curiosamente, los más lluviosos se encuentran en la parte oriental, como consecuencia derivada de poseer una mayor altitud, absoluta y relativa. Todo ello se traduce en que el coeficiente de correlación que se obtiene sea negativo, aunque con un porcentaje de explicación de la varianza bastante bajo, ya que en el mejor de los casos no

supera el 4%.

La exposición, por su parte, mantiene un coeficiente de correlación inferior al 25% con las precipitaciones medias durante buena parte del año, si bien esta situación cambia notablemente durante los meses estivales en los que apenas se llega al 2%. Esta relación, de carácter positivo, no debe sorprendernos, ya que se trata de uno de los factores geográficos que más inciden en las precipitaciones, dando lugar a un marcado incremento de las mismas.

Teniendo en cuenta que este es una análisis típico de una matriz de correlación, no nos cabe ninguna duda de que no resulta del todo adecuado. Esta falta de fiabilidad depende de una mala utilización de la técnica. Concretamente, se correlacionan variables de factor puro (factores geográficos -altura, emplazamiento, latitud, longitud y exposición-) con otras de factor impuro (las variables climáticas correspondientes -precipitaciones-).

De este modo, al aplicar la matriz de correlación lineal se produce un error metodológico, ya que se considera que las precipitaciones están modificadas por los factores geográficos. Esto es cierto, pero hay que considerar que las variaciones espaciales experimentadas por las precipitaciones se deben a la interacción de todos los factores geográficos, a la vez. Por lo tanto, no parece lógico aplicar una técnica bidimensional o bivariante, a un sistema complejo, multidimensional o multivariante.

Debido a esta circunstancia, proponemos la utilización de una técnica previa que permita "eliminar" la influencia supérflua que ejercen los factores geográficos que no intervienen en la matriz de correlación. Es decir, proponemos hacer una depuración de la influencia que ejerce cada factor geográfico en las precipitaciones y, seguidamente, establecer las correlaciones pertinentes entre cada factor geográfico y las modificaciones que provoca en las precipitaciones.

Para llevar a cabo este propósito proponemos la utilización de la regresión múltiple, de cinco variables independientes y tomando como variables dependientes a las precipitaciones de cada mes.

El procedimiento que seguimos es el siguiente:

1) Partimos de que las diferencias espaciales de la pluviometría dependen de los cinco factores geográficos señalados -altura, emplazamiento, latitud, longitud y exposición.

2) Si tomamos como variables dependientes a cuatro de ellos y calculamos los residuos, esto es, las diferencias entre el modelo real y el teórico. El resultado será que obtenemos la modificación, ya sea incremento o descenso,

experimentada por las precipitaciones en función del factor geográfico omitido en los cálculos.

Ejemplo:
Variables independientes: emplazamiento, latitud, longitud y exposición.
Variables dependientes: precipitaciones
Residuos: influencia que ejerce la altura, parámetro geográfico omitido.

3) Aplicamos la matriz de correlación lineal sobre los factores geográficos y la influencia que ejerce cada uno de ellos en las variables pluviométricas analizadas.

Con este procedimiento conseguimos aislar la influencia que ejerce cada factor geográfico en las precipitaciones, eliminando la variabilidad espacial provocada por la acción del resto de parámetros geográficos. De este modo, se aplica una técnica estadística bivariante a dos variables de factor puro, ya que las precipitaciones han sido sustituidas por la influencia que ejerce un elemento geográfico sobre ellas. Así, el fundamento metodológico de la matriz de correlación no experimenta ninguna irregularidad y, por ende, los resultados son mucho más convincentes, como puede comprobarse en la matriz de correlación simplificada que figura a continuación.

CUADRO 2

VAR. CLIMÁT.	ALTURA	EMPLAZ	LATITU	LONGIT	EXPOSI
I.F.G.P. EN.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. FB.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. MZ.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. AB.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. MY.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. JN.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. JL.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. AG.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. ST.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. OC.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. NV.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. DC.	0.910	0.670	0.700	0.950	1.000
I.F.G.P. AN.	0.910	0.670	0.700	0.950	1.000

En el cuadro precedente, obtenido mediante la aplicación del procedimiento expuesto, se observan algunas diferencias significativas con respecto al anterior.

En primer lugar, los coeficientes de correlación han aumentado de forma significativa, sobre todo en algunos factores geográficos, como consecuencia de haber depurado la información y, de esa forma, aplicar la técnica de la forma más idónea.

De este modo obtenemos que la correlación obtenida entre la influencia de los diferentes factores geográficos (I.F.G.) en las precipitaciones y dichos parámetros geográficos es muy elevada. El porcentaje de explicación de la varianza se incrementa bastante, sobre todo si lo comparamos con los obtenidos al efectuar la matriz de correlación entre variables de factor puro e impuro.

En segundo lugar, los coeficientes de correlación que hemos calculado son similares para cada mes del año y, además, para el caso anual. Esto implica un hecho ciertamente interesante, máxime si tenemos en cuenta que antes no aparecía. Nos referimos al comportamiento similar de los factores geográficos durante los diferentes meses del año. Esta circunstancia no debe extrañarnos ya que en condiciones de laboratorio se cumpliría de forma inexorable.

Por último, en tercer lugar, no aparecen las contradicciones físicas que obteníamos en el caso anterior. Es decir, nos encontramos con que la parte occidental de Extremadura tiene rasgos oceánicos, mientras que antes nos aparecían los rasgos oceánicos en la parte oriental, como consecuencia derivada de la ubicación de los observatorios seleccionados.

Teniendo en cuenta todos estos aspectos, llegamos a la conclusión de que el método propuesto elimina, al menos en climatología, buena parte de los riesgos de interpretación de la matriz de correlación lineal, con lo que si aplicamos otras técnicas complejas basadas en la misma, obtendremos unos resultados mucho más interesantes.

REFERENCIAS.

ANDERSON, T. W. (1984). *An introduction to multivariate statistical analysis*. Wiley. New York.

BASSIST, A. N. (1989). "The relationship between orography and precipitacion variability: a global view". En: *Sixth Conference on Applied Climatology*. Charleston.

CAPEL MOLINA, J. J. (1978). "Factores del clima en la Península Ibérica". *Paralelo 37º*, nº 4. Consejería de cultura. Junta de Andalucía. Excelentísima Diputación Provincial. Almería.

COMPÁN VÁZQUEZ, D. (1978). "Sobre el uso de la correlación lineal simple en Geografía". *Cuadernos Geográficos*, nº 8. Granada.

GURRÍA GASCÓN, J. L. (1984). "La Correlación Lineal: Precisiones prácticas y su funcionalidad en la determinación de las similitudes y diferencias de los espacios geográficos". *Norba V. Revista de Geografía*. Servicio de Publicaciones de la Universidad de Extremadura. Cáceres.

SÁNCHEZ MARTÍN, J. M. *El clima de montaña en Extremadura. Delimitación y análisis sistémico.* (Tesis Doctoral en elaboración).