SOBRE EL USO DE LA CORRELACION LINEAL SIMPLE EN GEOGRAFIA. APLICACION AL ESTUDIO DE LA DISTRIBUCION ESPACIAL DE LA RENTA EN ESPAÑA

DIEGO COMPAN VAZQUEZ

SUMMARY: In the first part of the article the correlation concept can be seen together with the different classes and the problems set by this technique applying it to investigation into regional phenomenons. Immediately afterwards the simple lineal correlation coefficient is obtained, step by step, by means of two procedures, the Pearson moment-product, and the gradual Spearman procedure. In the second part this same technique is applied to the study of the regional distribution of income per capita in Spain.

RESUME:
Dans le premiere partie de l'article on montre le concept de corrélation, ses différents types et la problematique qu'offre cette etechnique appliquée dans la recherche de phénomènes de portée espaciales. On obtient peu a peu a la suite, et por le même exemple, le coéfficient de corrélation linéaire simple porupour deux procédés celui du moment-produit de Pearson et le graduel de Spearman. Dans la deuxieme partie on applique dite technique a l'étude de la distribution spaciale du revenu par habitant en Espagne.

Este artículo pretende básicamente mostrar el concepto, significación, forma de obtención y problemática de la técnica de correlación lineal simple en Geografía (1). Está dirigido a todos los geógrafos y estudiantes de Geografía que desconocen esta técnica y parten de un nivel 0 en conocimientos matemáticos. No hemos tenido inconveniente de valernos de un ejemplo poco "espectacular" porque ello nos ha permitido extendernos en el comentario y dar una visión más completa de las posibilidades de la técnica.

En castellano, correlación significa analogía, relación recíproca entre dos o más cosas. En Geografía utilizamos este concepto con gran frecuencia. Cuando hablamos de la relación entre las precipitaciones y la altitud, o entre desarrollo e inversiones, no estamos sino manifestando la existencia de una correlación.

Las aseveraciones generales sobre la correlación entre dos hechos pueden tener gran valor si el geógrafo logra exponerlas con detalle. Pero si éste se apoya sólo en datos procedentes de elaboraciones matemáticas elementales, corre el riesgo de no dar una visión muy precisa, a la vez que, en muchos casos, no obtiene datos aptos para seguir desarrollándose en análisis más complejos, capaces de mostrar con mayor claridad las intrincadas situaciones que se presentan constantemente en Geografía.

Las técnicas estadísticas ofrecen un arsenal de instrumentos, cada día más numerosos y precisos, que están siendo desarrollados y, ya masivamente, utilizados en la geografía anglosajona. La correlación es una de las técnicas más simples y, también, una de las más útiles. La correalción simple puede describir con increible precisión el grado de asociación existente entre dos variables (2). Asi, aplicándola a un área geográfica concreta podríamos decir, por ejemplo, que la variable "altura sobre el nivel del mar" está asociada con la variable "precipitaciones medias anuales"; que esta relación tiene un sentido positivo (a mayor altitud, mayores

precipitaciones); una fuerza, medible en una escala que oscila entre 0,0 y 1,0, del orden, supongamos, de 0,74; y, finalmente, que la variable altitud conlleva, o implica, una serie de hechos que se distribuyen por el área de estudio de forma tal que "explican" el 55% ($0,74^2 \cdot 100$) de la variación experimentada por la variable precipitaciones a través de dicha superficie. El resto de la variación no explicada (45%) dependería de otros hechos "contenidos" en otras variables como podrían ser la orientación del terreno, la disposición general del relieve, la proximidad al mar, etc.

Pero no es sólo eso. Si desarrollamos un poco esta técnica, podríamos obtener una recta de regresión, con su fórmula correspondiente, que nos permitiría estudiar el problema gráficamente. A la vez, la regresión nos ayudaría mucho en el conocimiento de aquellas áreas específicas de la zona de estudio, donde el comportamiento de las precipitaciones, por ejemplo, se aparta de la tónica general. La regresión nos capacitaría incluso para predecir las precipitaciones en lugares donde sólo conocemos la altitud. La correlación propiamente dicha tiene otras posibilidades en Geografía ya que puede aplicarse al mismo tiempo a un número de variables muy superior a dos. Supongamos que en el ejemplo anterior hemos logrado medir conceptos como la cantidad de precipitaciones, la altitud, orientación y proximidad al mar en una serie de lugares distribuidos por toda el área de estudio. Entonces sería posible, por ejemplo, relacionar la primera variable con las otras tres al mismo tiempo y concluir, supongamos, que las tres últimas "explican" conjuntamente el 94% de la variación de las precipitaciones a través de toda la extensión superficial que estamos estudiando. Ahora tendríamos una idea mucho más precisa de los factores condicionantes de la lluvia, y tendríamos el camino allanado para estudiar las subáreas con características excepcionales, pero, además, valiéndonos de la regresión múltiple, podríamos averiguar, con escasa probabilidad de error, la cuantía de las precipitaciones de cualquier punto sin observatorio de la zona del que conocemos su altitud, distancia al mar y orientación. Un desarrollo más sofisticado de estas técnicas nos llevaría a otros tipos de análisis más complejos y perfeccionados, como el de la superficie de tendencia, el de los componentes principales o el factorial, capaces de aclarar sustancialmente las complicadas estructuras espaciales o funcionales de los hechos y situaciones que se presentan habitualmente en Geografía.

I. CLASES DE CORRELACION.

Existen diferentes clases de correlación que se pueden clasificar según criterios como el tipo de datos que necesitan, el carácter mismo de la relación en cuestión, o el número de variables puestas en relación.

Una correlación es *simple* o *bivariada* cuando sólo mide la realción existente entre dos variables, por ejemplo, la altitud y la pluviosidad de una zona. La correlación es *múltiple* o *multivariada* cuando mide la relación entre una variable y dos o más variables juntas, por ejemplo, entre la evolución de la renta per cápita de un país, a lo largo de una serie determinada de años, y otras variables como las inversiones en la industria y el saldo de la balanza de pagos durante la misma serie de años. La correlación es *canónica* cuando relaciona dos series de variables entre sí, por ejemplo, el desarrollo económico de una ciudad a lo largo de un periodo de treinta y cinco años (previamente medido por una serie de variables como renta per cápita, número de teléfonos, etc.) con el grado de accesibilidad de la misma (medido por otras variables como distancias en tiempo a otras áreas, frecuencia de autobueses, etc.) en el marco de una extensión superficial superior, como podría ser un área de mercado.

Para definir los restantes tipos de correlación vamos a referirnos solamente al caso de la correlación simple porque es la más fácil de

comprender y ello facilita notablemente la exposición.

Una correlación es *lineal* cuando al dibujar en un gráfico los valores que toman ambas variables, resultan formar, o tienden a formar, una línea recta (fig. 1—a y 1—b); en este caso, los valores que toman ambas variables varían, o tienden a variar, de una manera uniforme y constante. La correlación es *curvilínea* o *no lineal* cuando los valores de las variables, representados en un gráfico, forman, o tienden a formar, una línea curva (fig. 1—c); en este caso, las variables no varían de una forma constante y uniforme; un ejemplo sería la relación expresada por una función exponencial (3).

Una correlación es *positiva* o *directa* cuando la relación existente entre las variables tiene el mismo sentido, es decir, si una aumenta, la otra también, y viceversa (fig. 1—b y 1—c). En caso contrario, la correlación es *negativa* o *inversa* (fig. 1—a).

La correlación es *perfecta* cuando la variación de una variable se corresponde exactamente con la variación de la otra; por ejemplo, la existente entre la temperatura en grados centígrados y grados Farenheit para las mismas observaciones del mismo observatorio; si dibujamos los valores de ambas variables en un gráfico obtendríamos una línea (curva o recta) perfectamente definida (fig. 1—b). La correlación es *imperfecta* cuando la variación de una

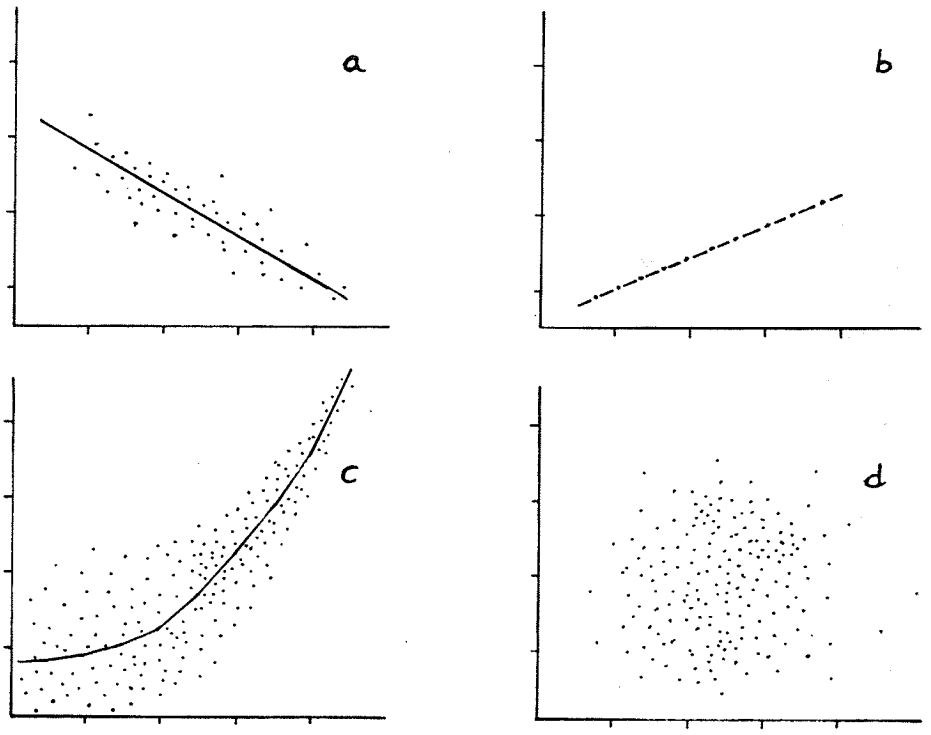

Fig. 1.- Diagramas d3 dispersión de valores correspondientes a dos variables y tipos de correlación resultantes (las líneas representan a las rectas de regresión, que indicarían la posición de los puntos en el caso en que la correlación fuese perfecta).

variable se corresponde sólo de forma parcial con la variación de la otra; dibujando los valores de ambas en un gráfico obtenemos un conjunto de puntos que tienden a formar una línea (figs. 1—a y 1—c). La correlación es *nula* o *inexistente* cuando ambas varían indistintamente sin guardar ningún tipo de relación; representados sus valores en un gráfico obtenemos una nube amorfa de puntos, no encajable en ninguna línea (fig. 1—d).

En Estadística, se llama correlación *sin sentido* aquella en que las variables relacionadas están a su vez influidas por otra variable común, que sería la que, en definitiva, explica esa correlación. En Geografía, este tipo de correlación puede tener importantes aplicaciones; así, una variable que no se puede medir, o de la que no disponemos de datos, puede ser sustituida en muchas ocasiones por otra variable conocida que esté fuertemente correlacionada con la primera por alguna causa común.

II. *LA CORRELACION SIMPLE.*

Es un método estadístico que mide numéricamente el grado de asociación, o la fuerza de la relación, que existe entre dos variables. Ello viene expresado mediante el valor de un coeficiente de correlación "r". Su obtención exige la previa definición de ambas variables por dos series de valores numéricos correspondientes a observaciones aparejadas. El coeficiente se somete luego a un test que nos indica, para ese caso concreto, la probabilidad estadística existente de que la asociación de ambas variables se deba puramente al azar. La correlación simple es, pues, una técnica estadística que nos dice estrictamente hasta qué punto la variación de una variable se refleja, o es reflejada por, la variación de la otra.

La correlación tiene un importante campo de acción en Geografía. Puede utilizarse básicamente en dos tipos de situaciones que aparecen sistemáticamente en el centro de todos los estudios de nuestra ciencia: A) Para analizar la evolución en el tiempo de dos hechos o características de un lugar o zona determinados. Por ejemplo, para conocer la relación existente entre la emigración y las tasas de inversión en una provincia durante los quince últimos años (cada par de valores estaría formado por los correspondientes al mismo año para ambas variables). B)Para conocer el grado de asociación existente entre dos hechos que varían a través de una misma superficie. Por ejemplo, para conocer la relación existente entre las precipitaciones de primavera y la cosecha de trigo de secano en las diferentes comarcas de Castilla la Nueva (cada par de valores estaría formado por los datos correspondientes a un mismo lugar para ambas variables).

El coeficiente de correlación mide el grado de asociación existente entre las variaciones experimentadas por ambas variables. Este, independientemente del tipo de datos de que se parte, siempre está comprendido entre los valores +1,0 y —1,0, así pues, no sólo mide la intensidad o la fuerza de la correlación entre ambas variables, sino, también, el sentido de la misma.

Si "r" es menor que 0, la correlación es negativa o inversa (fig. 1—a). Si "r" es mayor que 0, la correlación será positiva o directa (figs. 1—b y 1—c). Si "r" es igual a 0, habrá incorrelación o correlación nula (fig. 1—d). Si "r" es igual a 1,0 tendremos una correlación perfecta (fig. 1—b).

Del coeficiente de correlación "r", que mide el grado de asociación entre las variables, se obtiene el *coeficiente de determinación* "r^2", que mide la cantidad de variación de una variable que está "explicada" por, o "contenida" en, la variación de la otra. Esta cantidad de explicación suele darse en forma de porcentaje multiplicando "r^2" por 100.

El coeficiente de correlación "r" mide sólo el sentido y la fuerza de la relación, pero no muestra si esta relación es o no causal. Un alto

coeficiente de correlación puede deberse, en efecto, a varios motivos. Puede haber una relación causa-efecto. Puede haber una causa común que condicione a ambas variables (ésta podría incidir con fuerza diferente en cada una de ellas; a la vez, ambas variables podrían estar condicionadas por otras variables independientemente). Finalmente, la relación puede deberse al azar. Todo ello quiere decir que la técnica no provee de una explicación de la realidad y que el juicio crítico del geógrafo es absolutamente necesario.

Entre los diferentes coeficientes de correlación lineal simple que existen, aquí sólo vamos a tratar con dos, el *coeficiente de correlación r del momento-producto de Pearson* y el *coeficiente de correlación gradual, o del rango* r_s *de Spearman.*

Ambos coeficientes se basan en la idea de la covarianza, que describe estadísticamente la correspondencia entre la variación de dos variables (4). Para obtener la covarianza de dos variables se busca la desviación sobre la media correspondiente de cada uno de los n valores de que se componen las variables, esto es: $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, ... $(x_n - \bar{x})$ para la variable X, y $(y_1 - \bar{y})$ $(y_2 - \bar{y})$ ... $(y_n - \bar{y})$ para la variable Y; donde $\bar{x}$ e $\bar{y}$ son los valores medios de ambas variables, y $(x_1, x_2, \ldots, x_n)$ e $(y_1, y_2, \ldots, y_n)$ representan respectivamente a x_i e y_i, es decir, cada uno de los n valores de que se compone cada variable. A continuación se multiplican las desviaciones correspondientes a cada par de valores ($(x_1 - \bar{x}) \cdot (y_1 - \bar{y})$; $(x_2 - \bar{x}) \cdot (y_2 - \bar{y})$, etc.). La suma de todos los productos (a los resultados con signo positivo se restan los de signo negativo) nos dará la desviación total. Esta última se divide por el número de pares, n, y se obtendrá la desviación promedio, que no es otra cosa que la covarianza. Su fórmula es:

$$\text{Covarianza} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} \qquad (F\text{-}1)$$

Donde $\sum_{i=1}^{n}$ es el signo sumatorio; indica la suma de todos los valores de i cuando i varía entre "1" y "n" (en este caso sería la suma de todos los productos correspondientes a cada par de desviaciones).

Si la covarianza es grande y positiva, la variación de las variables y las varianzas de éstas también lo serán. La covarianza es una medida que depende de la magnitud de los datos en que aparece medida cada variable. Por ello tenemos que estandarizarla si queremos que sea una medida universalmente comparable. Ello se consigue dividiendo la covarianza por el producto de las desviaciones estándar (5) de las dos variables. El coeficiente de correlación "r" no es otra cosa que la covarianza estandarizada. Su fórmula es:

$$r = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y} \qquad (F\text{-}2)$$

II. *EL COEFICIENTE DE CORRELACION "r" MOMENTO_PRODUCTO DE PEARSON.*

Es el coeficiente de correlación más preciso y útil de cuantos existen. En él descansan otras técnicas estadísticas más desarrolladas, como el análisis factorial, que se basan en el concepto de la correlación. Su fórmula es exactamente la que vimos para la covarianza estandarizada (F—2). También se utiliza en otra versión que ahorra tiempo en el cálculo cuando sólo se dispone de una calculadora manual simple:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}} \qquad (F\text{-}3)$$

Este coeficiente se basa en una serie de presupuestos estadísticos muy exigentes que pueden restringir notablemente su uso. El coeficiente "r" necesita que las variables estén medidas

con valores en la escala de intervalo (6), que la relación entre las variables sea de tipo lineal (7), y que los elementos o valores que integran las variables provengan de una distribución de tipo normal (8).

La obtención del coeficiente de correlación implica, cuando menos, cinco etapas: 1.ª, selección de las variables que queremos estudiar, 2.ª, elección de la forma de cuantificar las variables, de modo que expresen adecuadamente las ideas contenidas en las mismas. 3.ª, sometimiento de los datos a algunas pruebas que nos indiquen si cumplen o no los requisitos estadísticos que hemos mencionado más arriba. 4.ª, obtención del coeficiente mediante la aplicación de la fórmula (F—2), o (F—3). 5.ª, sometimiento del coeficiente obtenido a una prueba o test que nos indique el grado de probabilidad existente de que dicha correlación no sea real y se deba solamente al azar.

IV. *EL COEFICIENTE DE CORRELACION GRADUAL, O DEL RANGO,* r_s *DE SPEARMAN.*

Este coeficiente se basa en la misma idea que el coeficiente r del momento-producto. Se diferencia de éste en que utiliza datos en la escala ordinal (ver nota 6) y en que los datos utilizados no necesitan venir de una distribución normal.

Aunque es menos exacto que el del momento-producto, se aconseja usarlo cuando los datos en la escala de intervalo no sean de fiar, pero sí lo sea el orden que éstos ocupan en la escala ordinal; cuando los valores de alguna variable no provengan de una distribución normal; o, simplemente, cuando sólo conocemos los datos en la escala ordinal. Este coeficiente es más rápido de obtener que el del momento-producto, por ello puede obtenerse como alternativa rápida en los casos en que sólo estamos interesados en tener una idea aproximada de la relación entre dos variables, por ejemplo, cuando estamos haciendo tanteos.

El coeficiente de correlación gradual r_s no da una medida totalmente eficiente de la correlación ya que no se basa en el valor de la observación —que sería apreciado en la escala de intervalo—, sino en la posición que éste ocupa en una escala ordenada. Hammond & McCullagh nos dicen que este coeficiente sólo tiene una eficacia potencial del orden del 91% de la del coeficiente de correlación r del momento-producto (9).

Otra limitación relativamente importante de este coeficiente es que cuando se utiliza para menos de 10 pares de observaciones o valores, o, para más de 30, puede contener un grado de error importante (10).

La fórmula de este coeficiente es:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (F\text{-}4)$$

Donde n es el número de pares de valores u observaciones y d_i es la diferencia entre el número de orden que cada uno de los i valores ocupa en ambas variables.

V. *EJEMPLO DE OBTENCION DE LOS COEFICIENTES DE CORRELACION "r" y "*r_s*".*

A continuación vamos a obtener ambos coeficientes de una forma práctica, basándonos en un mismo ejemplo. Como ya hemos visto, la correlación simple necesita dos series de datos apareados que tienen en común el tiempo o el espacio, y se basa en la idea de la comparación de ambas para saber si los cambios o variaciones de una de ellas están o no reflejados en los cambios o variaciones de la otra.

Supongamos que estamos tratando de averiguar cuál es la relación existente entre la

variación de la renta per cápita en cada una de las partes del país, y la variación de la proporción interna de trabajadores en cada una de esas partes. En términos estadísticos se trataría de conocer cuál es la relación entre la variación experimentada por ambas variables a través del espacio nacional. En este caso, si a igual tiempo de trabajo correspondiese igual salario o idénticas ganancias, deberíamos obtener una correlación casi perfecta (11), que nos indicaría que la hora de trabajo se paga igual en cualquier parte del país. Esto, obviamente, no es así, y la correlación nos brinda la ocasión de medirlo de forma numérica.

Una vez seleccionadas las dos variables y la magnitud en que se van a medir (renta per cápita y porcentaje de población activa), hemos de seleccionar las unidades de muestreo más adecuadas. En la selección de la unidad superficial de base, el geógrafo cuenta con un obstáculo de primera magnitud, los datos de que suele disponer se refieren a unidades de computación estadística como el municipio, la provincia, la región o el país. Mientras tanto, en la ciencia estadística, la correlación se basa en la manipulación de una muestra representativa de la variable en cuestión, formada por diferentes observaciones al azar en diferentes unidades de muestreo de idénticas características (idénticas dimensiones en nuestro caso). El problema de la correlación, aplicado a datos de contenido espacial, estriba básicamente en que, a pesar de utilizarse medidas superficiales relativas (densidades, etc.), los resultados de la correlación suelen variar un poco al modificar las dimensiones de la unidad superficial de base: "diferentes escalas pueden producir diferentes grados de relación... cuando las unidades de superficie aumentan de extensión, la variación de la variable dependiente disminuye, y aumenta el coeficiente de correlación" (12).

Lo ideal sería tener dividido el país en pequeñas unidades de idénticas dimensiones (por ejemplo, unidades de 1 km , que ya existen como unidades de computación estadística, para algunos fines, en paises como Suecia o el Reino Unido) y disponer de datos para cada uno de los cuadros. En este caso, para nuestro estudio bastaría manejar una muestra representativa de cuadrículas de las distintas partes del país. Ante este problema se han tomado algunas soluciones, no generalmente satisfactorias, como el "pesado" o valoración relativa de cada superficie según ciertos criterios (13), y los sociólogos y estadísticos han abierto un controvertido debate sobre el tema, por el momento sin solución, que en los paises anglosajones se conoce como la *falacia ecológica*.

De cualquier forma, a pesar de estas limitaciones, las técnicas de correlación, aplicadas a unidades espaciales, han resultado ser tan útiles en Geografía que, incluso transgrediendo parcialmente los principios de la ciencia estadística en que se basan, se utilizan masivamente en los paises anglosajones. Con ellas, en efecto, se obtienen unos resultados mucho más positivos que si se utilizaran los caminos de la Geografía tradicional. En este sentido, Taylor justifica el uso de la correlación en Geografía desde este punto de vista: "El criterio para la utilización de un modelo no es el de la bondad de su actuación estadística, sino su utilidad geográfica", y concluye en que el problema es sólo de tipo metodológico: "Tomamos prestada la teoría de otras ciencias sociales. Es decir, tomamos un modelo aespacial y lo adaptamos para un contexto espacial... El principal problema de la asociación superficial es metodológico, es la falta de un armazón teórico"(14).

Siguiendo con nuestro ejemplo, podríamos tomar una muestra de municipios distribuidos por el territorio nacional, pero, por falta de datos publicados a este nivel, nos decidimos por unidades superiores. La región podria ser una unidad adecuada, pero nos daría una visión poco detallada y pobre en matices. Por ello, nos decidimos por la unidad superficial de la provincia (15). Luego, habría que pensar si

tomar sólo una muestra de provincias o incluirlas todas; ante el riesgo de error que implicaría una muestra y el escaso exceso de trabajo que supondría operar con todas, nos decidimos por la segunda posibilidad (16).

Ya tenemos pues, las dos variables ("renta per cápita media provincial" y "porcentaje de población activa interna provincial"). Ambas variables están integradas por 50 pares de valores (datos de ambas variables referidos a las 50 provincias). Seguidamente, antes de someter los datos a la fórmula de la correlación, tendríamos que asegurarnos de que los datos de ambas variables proceden de una población estadística distribuida "normalmente". La forma más rápida es dibujar dichos datos en sendos histogramas de frecuencias, que nos dan una visualización rápida, aunque aproximada, del problema (ver fig. 2). El histograma correspondiente a la renta per cápita nos indica que se trata, a grandes rasgos, de una serie de valores distribuidos de una forma aproximadamente "normal". La curva de frecuencias indica, no obstante, una tendencia a desviarse hacia la izquierda (desviación positiva en términos estadísticos) y la existencia de un pequeño máximo secundario en el extremo derecho. Ello se debe a que hay unas cuantas provincias que gozan de una renta anormalmente alta (Vizcaya, Barcelona, Madrid, Guipúzcoa, Alava). Nosotros podríamos pensar, por el momento, que se trata de una distribución aceptablemente normal, susceptible de ser utilizada en la correlación, pues, sobre todo, hemos de tener en cuenta que nuestros datos deben estar sujetos, indudablemente, a cierto error de medida (la unidad superficial aceptada, la provincia, es sólo una de las infinitas unidades superficiales posibles (17). De cualquier forma, puesto que existen pruebas estadísticas para aceptar o rechazar a una distribución como "normal", vamos a someter a nuestros datos a una de las más simples. Siguiendo el procedimiento indicado en la nota 8, hemos dividido a nuestros datos estandarizados (ver cuadro 1, "X estandarizada") en cuatro grupos separados por los niveles de desviación estandar de +0,67; 0,0 y —0,67. En dichos grupos aparecen, respectivamente, 10,11,15 y 14 valores provinciales (si se tratase de una distribución perfectamente "normal", en cada grupo debería de haber 12,5 valores, es decir, la cuarta parte del total. Aplicando la fórmula F—10 obtendremos un valor X del 1,36; esdecir, un valor muy inferior a 2,71, que

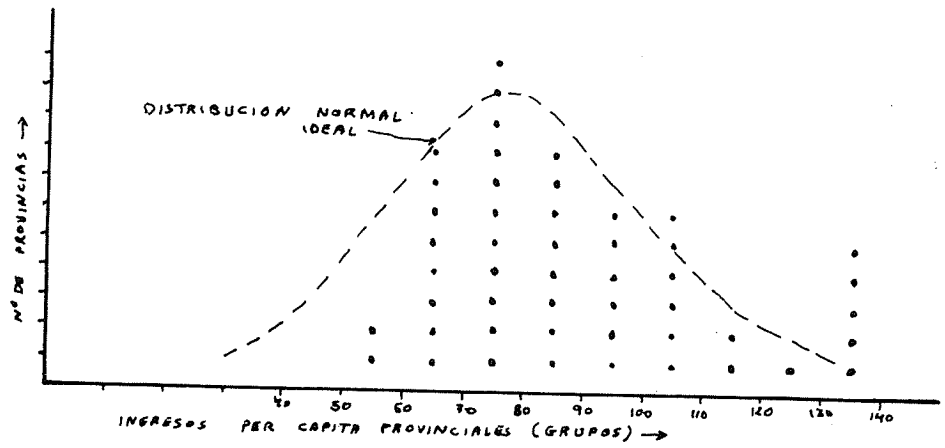

Fig. 2.- Diagrama de frecuencias de variables "renta per cápita" media provincial, que incluye valores para las cincuenta provincias.

OBTENCION DE LOS COEFICIENTES DE CORRELACION "R" y "R_s" entre las variables renta per cápita provincial y tasa de actividad provincial.

	(Miles ptas.) ingresos provinciales por hab. x	*% Población activa sobre total prov.* y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	*Número de orden decreciente* x	y	*(Diferencia)* d	d^2	*x estandarizada*
Alava	131,1	40,3	42,51	1,70	72,27	4	9	-5	25	1,96
Albacete	68,0	36,3	-20,59	-2,30	47,36	42	35	7	49	-0,95
Alicante	96,3	38,8	7,71	0,20	1,54	17	22	-5	25	0,35
Almería	71,0	33,4	-17,59	-5,20	91,47	40	49	-9	81	-0,81
Avila	64,2	35,8	-24,39	-2,80	68,29	45	39	6	36	-1,12
Badajoz	60,9	35,6	-27,69	-3,00	83,07	48	41	7	49	-1,28
Baleares	120,0	40,1	31,41	1,50	47,11	6	14	-8	64	1,45
Barcelona	132,4	40,2	43,81	1,60	70,10	3	12	-9	81	2,02
Burgos	88,0	39,1	-0,59	0,50	-0,30	22	19	3	9	-0,03
Cáceres	58,8	37,0	-29,79	-1,60	47,66	49	31	18	324	-1,37
Cádiz	74,1	32,5	-14,49	-6,10	88,39	37	50	-13	169	-0,67
Castellón	95,3	40,7	6,71	2,10	14,09	18	8	10	100	0,31
C. Real	74,2	37,1	-14,39	-1,50	21,58	35	30	5	25	-0,66
Córdoba	68,7	35,5	-19,89	-3,10	61,66	41	43	-2	4	-0,92
L. Coruña	77,6	45,5	-10,99	6,90	75,83	33	4	29	841	-0,51
Cuenca	74,2	36,6	-14,39	-2,00	28,78	36	33	3	9	-0,66
Gerona	117,8	42,5	29,21	3,90	113,92	7	5	2	4	1,35
Granada	62,7	33,7	-25,89	-4,90	126,86	47	48	-1	1	-1,19
Guadalajara	92,3	37,3	3,71	-1,30	4,82	20	29	-9	81	0,17
Guipúzcoa	130,4	39,7	42,06	1,10	46,26	5	15	-10	100	1,93
Huelva	72,3	35,5	-16,29	-3,10	50,50	39	44	-5	25	-0,75
Huesca	96,8	39,3	8,21	0,70	5,75	16	18	-2	4	0,38
Jaén	64,5	36,0	-24,09	-2,60	62,63	44	36	8	64	-1,11
León	79,5	40,2	-9,09	1,60	14,54	29	10	19	361	-0,42
Lérida	110,6	39,6	22,01	1,00	22,01	8	17	-9	81	1,02
Logroño	100,9	41,3	12,31	2,70	33,24	14	6	8	64	0,57
Lugo	57,6	52,4	-30,99	13,80	-427,66	50	1	49	2401	-1,43
Madrid	138,6	38,4	50,01	-0,20	-10,00	2	24	-22	484	2,31
Málaga	76,5	35,7	-12,09	-2,90	35,06	34	40	-6	36	-0,56
Murcia	78,2	36,9	-10,39	-1,70	17,66	31	32	-1	1	-0,48
Navarra	109,9	38,6	21,31	-0,00	-0,01	9	23	-14	196	0,98
Orense	62,9	51,6	-25,69	13,00	333,97	46	2	44	1936	-1,18
Oviedo	94,4	40,2	5,81	1,60	9,30	19	11	8	64	0,27
Palencia	77,8	35,4	-10,79	-3,20	34,53	32	45	-13	169	-0,50
L. Palmas	86,3	34,9	-2,29	-3,70	8,47	23	46	-23	529	-0,11
Pontevedra	83,1	47,8	-5,49	9,20	-50,51	24	3	21	441	-0,25
Salamanca	72,9	35,6	-15,69	-3,00	47,07	38	42	-4	16	-0,72
S. C. Tenerife	80,0	34,7	-8,59	-3,90	33,50	27	47	-20	400	-0,40
Santander	102,2	39,6	13,61	1,00	13,61	13	16	-3	9	0,63
Segovia	82,9	36,5	-5,69	-2,10	11,95	25	34	-9	81	-0,26
Sevilla	79,0	35,9	-9,59	-2,70	25,89	30	38	-8	64	-0,44
Soria	88,9	38,1	0,31	-0,50	-0,50	21	26	-5	25	0,01
Tarragona	107,0	40,1	18,41	1,50	27,61	10	13	-3	9	0,85
Teruel	79,8	38,9	-8,79	0,30	-2,64	28	20	8	64	-0,41
Toledo	81,1	38,8	-7,49	0,20	-1,50	26	21	5	25	-0,34
Valencia	102,8	38,3	14,21	-0,30	-4,26	11	25	-14	196	0,65
Valladolid	99,1	36,0	10,51	-2,60	-27,33	15	37	-22	484	0,48
Vizcaya	138,8	37,9	50,21	-0,70	-35,15	1	27	-26	676	2,32
Zamora	65,3	40,8	-23,29	2,20	-51,24	43	7	36	1296	-1,07
Zaragoza	102,4	37,4	13,81	-1,20	-16,57	12	28	-16	256	0,64

$\bar{X} = 88,59$; $\sigma_x = 21,60814$; $\sum_{i=1}^{n} (x_i-\bar{x})\cdot(y_i-\bar{y}) = 1469,19 - 1056,48 = 412,71$

$\bar{Y} = 38,60$; $\sigma_y = 3,92354$; $\sum_{i=1}^{n} d_i^2 = 12534$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{50}\cdot 412,71}{85,06289} = +0,097 \; ; \quad r^2 = 0,01$$

$$r_s = 1 - \frac{6\cdot\sum_{i=1}^{n} d_i^2}{n^3-n} = 1 - \frac{6\cdot 12534}{124950} = 1 - 0,60 = 0,398 \; ; \quad r_s^2 = 0,158$$

Fuente: Los valores de ambas variables han sido obtenidos en Banco de Bilbao: "La renta nacional y su distribución provincial en 1973"

nos permite aceptar como válida la hipótesis nula, que, en este caso, establece que ambas distribuciones son idénticas.

En este momento ya estamos en condiciones de aplicar la fórmula de la correlación a nuestra serie de datos aparejados.

V.a. Obtención del coeficiente de correlación "r" del momento-producto de Pearson.

La forma más aconsejable para su obtención es construyendo una tabla como la que aparece en el cuadro 1 y aplicando la fórmula F—2. Sustituyendo:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{50} \cdot 412{,}71}{85{,}06} = 0{,}097 = 0{,}1$$

Obtenemos un valor de "r" = 0,1. Antes de comentar el significado del coeficiente convendría someterlo a una prueba de significación que nos indique cuál es la probabilidad estadística de que esta correlación no se deba a una relación real, sino sólo al azar. Antes de hacer esto, conviene aclarar que en este caso concreto no hace falta en absoluto someter nuestro coeficiente a la mencionada prueba de significación. Esta sería necesaria cuando, de acuerdo con los postulados estadísticos, las variables se obtienen a partir de una muestra al azar. En nuestro ejemplo no cabe duda de que el coeficiente es totalmente representativo ya que, mal que bien, hemos incluido la totalidad de los datos nacionales, sintetizados en unidades superficiales. No obstante, puesto que en muchos casos el geógrafo tendrá que basarse sólo en unas cuantas muestras al azar, vamos a exponer seguidamente el mecanismo de la prueba.

Todas las pruebas de significación, como los demás "tests" estadísticos, se basan en la idea de la aceptación o rechazo de una hipótesis nula o de su alternativa. La hipótesis nula postula, en este caso, que el coeficiente de correlación es igual a cero, es decir, postula que no existe correlación. Si esto no se demuestra ser cierto, tendríamos que aceptar la hipótesis alternativa, es decir, la que postula que sí hay correlación. Para la correlación, la prueba más usual es la *prueba t de Student*, que obtiene un índice "t" mediante la aplicación de la fórmula:

$$t = \frac{r\sqrt{n-2}}{1 - r^2} \qquad (F\text{-}5)$$

dónde n—2 es el número de observaciones "n" menos 2; el resultado de la sustracción se conoce como *número de grados de libertad* del caso o muestra en cuestión. "r" es el coeficiente de correlación. Este índice "t", con sus grados de libertad correspondientes, se contrasta posteriormente en una tabla de *valores críticos*, que suelen traer todos los manuales de Estadística general, y así podríamos conocer el grado de probabilidad de que el coeficiente se debiera al azar. Si la probabilidad es ínfima (por debajo del nivel de confianza elegido), se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Normalmente, en Estadística suelen utilizarse básicamente tres "límites de confianza" para la aceptación o rechazo de hipótesis: el de 0,05; 0,01 y 0,001; la elección de uno u otro de ellos vendría determinada por la necesidad de precisión que tengamos en cada caso. El primer límite es poco exigente, indica que a partir de él hay una probabilidad del 5%, o más, de que el coeficiente obtenido se deba al azar (en todo caso, este grado de probabilidad suele considerarse suficientemente elevado como para no rechazar la hipótesis nula). El segundo límite indicaría una probabilidad del orden del 1%; normalmente, con una probabilidad tan pequeña como esta, se rechaza la hipótesis nula y se acepta la alternativa. El tercer límite indica una probabilidad del orden del 0,1%; este límite es tan seguro que se utiliza para los casos más delicados solamente.

Aquí, para aligerar, más que incluir la mencionada tabla, incluimos un gráfico obtenido por Gregory (18) que se basa en la mencionada prueba (fig. 3). En él se puede relacionar directamente el coeficiente obtenido con los grados de libertad correspondientes, y se obtiene un punto en el gráfico que muestra visualmente la posibilidad de "error". Así, en nuestro ejemplo, un coeficiente de correlación "r" = 0,1, para un caso con 48 grados de libertad, producirá un punto "A", situado muy por debajo de la curva correspondiente al límite de confianza del 5%; con ello, no podríamos rechazar la hipótesis nula, esto es, habría más de un 5% de probabilidad de que el coeficiente obtenido se debiera al azar, y no estaríamos "autorizados" a decir que hay una correlación real entre ambas variables (19). En realidad, repetimos, en nuestro ejemplo no necesitábamos someter el coeficiente a la mencionada prueba; si lo hemos hecho ha sido para mostrar el mecanismo de funcionamiento de este tipo de "tests", tan necesarios en Estadística, y el hecho de que estos no son útiles cuando se trata de coeficientes muy próximos a cero.

Veamos ahora qué quiere decir un coeficiente de correlación de 0,1. Normalmente se acepta entre los estadísticos el valor de 0,5 para definir el límite entre lo que sería un coeficiente significativo o no: "En términos generales, son bastante significativos los coeficientes comprendidos entre +0,5 y +1,0; y entre —0,5 y —1,0; mientras que si los valores están comprendidos entre +0,5 y —0,5, hay que esperar una correlación poco significativa" (20). En nuestro ejemplo, un coeficiente de 0,1 nos sirve para rechazar totalmente la hipótesis de que en España existe una correlación clara y directa entre la proporción de trabajadores en cada provincia y la renta per cápita; es decir, que la variación existente en la primera variable a lo largo de la superficie nacional, no se refleja prácticamente en la variación correspondiente a la segunda variable; o, lo que es lo mismo, que ambas variables varían de una forma totalmente independiente. Un coeficiente de correlación "r" de 0,1, implica un coeficiente de determinación "r^2" del orden del 0,01, es decir, que sólo el 1% (0,01 x 100) de la variación espacial de la renta per cápita pro-

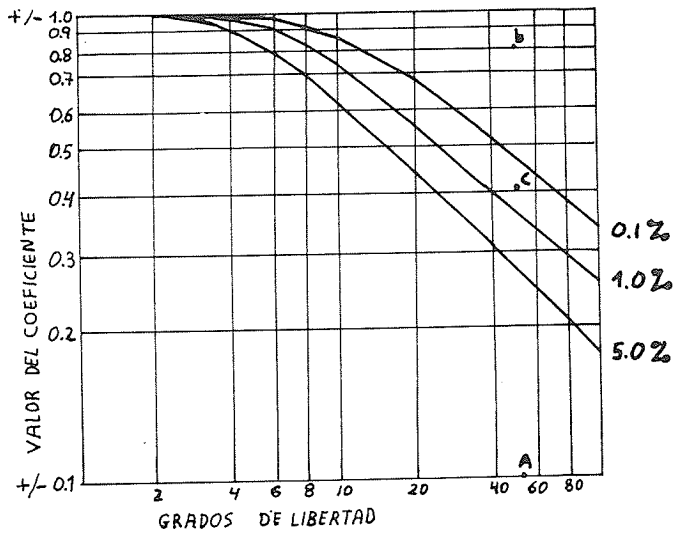

Fig. 3.- Gráfico de los niveles de significación de los coeficientes de correlación, utilizando la distribución *t* de Student.
(Fuente: Gregory, 1975, pág. 198).

vincial está "explicado" por la variación de la variable tasa de actividad provincial a través de la misma superficie nacional, y viceversa. En nuestro ejemplo tenemos dos provincias con características excepcionales: Lugo y Orense, que, teniendo las rentas más bajas del país, cuentan con las tasas de actividad más elevadas (ver cuadro 1). De esta forma, habría para ellas una fortísima correlación negativa entre las dos variables, y ello es así de forma tan intensa que la covarianza total nacional de las dos variables ($\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})$) sería, aproximadamente, el doble de la que es, si estas dos provincias tuviesen unas características medias, es decir, si sus valores correspondientes de $(x-\bar{x})$ ó $(y-\bar{y})$ fuesen iguales a cero (ver cuadro 1). Pensando, incluso, que este caso excepcional pudiese deberse a un error en la forma de obtención de los datos por el INE, hemos obtenido un nuevo coeficiente de correlación para las mismas variables, referidas al total de las provincias, suponiendo que ambas provincias gallegas suman cero a las varianzas de las dos variables; pero aún así, sólo se obtiene un coeficiente de correlación "r" de 0,28, y un coeficiente de determinación "r^2" de 0,08, es decir, unos valores suficientemente pequeños aún como para poder seguir manteniendo que la correlación entre ambas variables es extremadamente débil.

Si aún tenemos que explicar lo que significa este coeficiente, diremos que en España hay una debilísima relación positiva entre la distribución espacial de la renta per cápita y la distribución relativa de la tasa de actividad a través de la misma superficie. En otras palabras, que en nuestro país se dan tales condiciones que la renta per cápita obtenida en cada provincia (su promedio) no tiene nada que ver con el porcentaje de población que trabaja en cada una de ellas.

Un razonamiento más exigente nos llevaría, probablemente, a replantearnos cuestiones a varios niveles. Por un lado, tendríamos que pensar si hemos infringido seriamente los postulados estadísticos en que se basa la técnica, no parece, ciertamente, que ello haya ocurrido así. Sin embargo, lo que sí podemos ver es que la varianza de la variable "tasa de actividad provincial" es muy pequeña. Suficientemente pequeña como para necesitar ser medida con toda precisión; cuando hay diferencias de escasa cuantía entre los valores, cualquier error de medida puede conducirnos a resultados notoriamente erróneos. En nuestro caso, la variable "tasa de actividad" tiene una desviación estándar de sólo 3,9 para un valor medio de 38,6 y una serie de 50 valores, es decir, dicha variable tiene una desviación sobre la media suficientemente pequeña como para exigir una medida muy precisa. Aquí tendríamos que plantearnos de nuevo si los datos que hemos utilizado para medir dicha variable son de fiar; en este sentido, nuestros datos sí que parecen ofrecer una dosis considerable de duda, y ello ha podido conducirnos a un resultado no excesivamente real: los criterios seguidos por el INE para clasificar la población activa en el sector primario, parecen variar notablemente de una región a otra; no se podría decir, como parece deducirse de los datos (ver Cuadro 1) que en la agricultura de Orense trabajen casi todas las mujeres de la provincia, mientras que en la agricultura de Jaén no trabaje ninguna. De esta forma, podríamos decir que mientras no existan datos más precisos no será posible conocer, al menos con cierto detalle, hasta qué punto la distribución espacial de las tasas de actividad se refleja en la distribución espacial de la renta per cápita en nuestro país. No obstante, los resultados de la correlación siguen teniendo cierta validez, al menos como para poder decir que, muy probablemente, existe una relación positiva entre ambas variables, y que esta relación es bastante débil.

V.b. Obtención del coeficiente de correlación gradual "r_s" de Spearman.

Este coeficiente se obtiene de una forma mucho menos complicada que el del momento-producto. Para obtenerlo tenemos que asignar a

cada observación un valor que será igual al número de orden que ocupa ésta en una escala que previamente se ha ordenado de mayor a menor. Así, por ejemplo, a Vizcaya, con la renta per cápita más alta del país, habría que asignarle el valor "1" en la variable "renta", mientras que en la variable "tasa de actividad" habría que darle el valor 27, por ser la 27ª provincia española según el porcentaje de su población activa. Con las restantes provincias tendríamos que aplicar el mismo criterio de valoración.

Una vez computadas todas las observaciones, los valores de ambas variables se anotan formando pares, tal como aparece en el cuadro 1. Luego se obtiene la diferencia "d" de los valores de cada par. Posteriormente se eleva al cuadrado cada una de las diferencias "d", y se suman todos los resultados ($\sum_{i=1}^{n} d_i^2$). En ese momento ya tenemos todos los datos que requiere la fórmula del coeficiente de correlación gradual:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 12.534}{12.500 - 50} = 1 - 0{,}6019 = 0{,}398$$

Para conocer el grado de probabilidad de que el coeficiente se pueda deber al azar, habría que aplicar el mismo método expuesto para el coeficiente de correlación del momento-producto (en nuestro caso particular, como vimos, no es necesario obtener el coeficiente a la mencionada prueba de significación). El valor " r_s " podría ser contrastado en el mismo gráfico de la figura 3. En nuestro ejemplo, un coeficiente " r_s " de 0,4, para un caso con 48 grados de libertad, produce un punto "C" que se sitúa entre los límites de confianza del 1,0% y 0,1% (ver figura 3), lo que nos indica que existe menos del 1% y más del 0,1% de probabilidad de que el coeficiente se deba a un error. O, lo que es lo mismo, que existe más del 99,0% de probabilidad de que el coeficiente obtenido responda a una correlación real.

Aplicando ambos métodos al mismo ejemplo, hemos obtenido unos coeficientes de correlación bastante diferentes: $r = 0{,}1$, y $r_s = 0{,}4$. Normalmente, la diferencia entre ambos coeficientes, debida a la mayor inexactitud del segundo procedimiento, suele ser muy pequeña cuando la aplicamos al mismo caso (por ejemplo, 0,82 y 0,80 ó 0,15 y 0,12). Aquí hemos obtenido una diferencia excepcionalmente elevada. Ello se debe en parte a la insuficiencia de los datos de medida de la variable tasas de actividad (en nuestro ejemplo hay varios grupos de provincias en que la variación de la mencionada tasa es sólo del 0,1%; con ello, se multiplican los efectos de un posible error de medida, sobre todo en el caso del coeficiente de correlación gradual). Además, existe otra razón para esta notable diferencia entre los coeficientes: el coeficiente " r_s " debe tener aquí un grado de error bastante importante aquí porque ha sido obtenido a partir de 50 valores, cifra muy superior a la de 30 valores, que habíamos considerado como límite máximo para que el coeficiente produjese resultados aceptables. En cualquier caso, aún siendo " r_s " muy superior a "r", su valor, 0,4, sigue siendo inferior al que se suele considerar como de mínima significación estadística ($r = 0{,}5$). Siendo r_s^2 de sólo 0,16, podemos decir que la variación de cualquiera de dos variables no "explica" más que el 16% de la otra, es decir, una parte insignificante.

V.C. A modo de conclusión.

Así, a través de ambos coeficientes hemos llegado a la misma conclusión: en España no existe una relación significativa entre la cantidad de gente que trabaja (porcentaje de población activa provincial) y la renta per cápita que se percibe (media provincial). Por tanto, aún tendríamos que seguir investigando si lo que tratamos es de encontrar las razones que condicionan la desigual distribución de la renta per cápita en nuestro país. Esta tarea es fácil de resolver si tenemos en cuenta que, en España, debido al notorio desarrollo del "mecanismo de

intercambio desigual", los salarios o beneficios económicos dependen mucho más del tipo de trabajo que se hace que de la cantidad del mismo (en 1973, por ejemplo, según datos del Banco de Bilbao, los agricultores obtuvieron un promedio de 117000 pts. al año, mientras que los trabajadores en los servicios obtuvieron unas 237000 pts.). Basándonos en esto, y a la vista de que la descomposición sectorial de la actividad no es uniforme a través de la superficie nacional, hemos confeccionado el cuadro 2, donde se correlaciona (técnica del momento-producto) la variable "renta" con otras variables representativas de las tasas de actividad en los distintos sectores. Los resultados del cuadro pueden sorprender un poco, sobre todo en lo que respecta al sector terciario; siendo el mayor del país, tanto por el porcentaje de población que trabaja en el —38,4%—, como los beneficios por trabajador, su variación a través de la superficie nacional está poco relacionada con la variación de la renta per cápita; a penas puede "explicar" el 17% de la variación espacial de la renta. Este resultado es, ciertamente, lógico porque la variable "tasa de actividad en los servicios" no es uniforme en sí misma, engloba diferentes subsectores de muy diversa índole y significación económica que no se distribuyen uniformemente por el país en relación con la importancia que el sector tiene en cada provincia. Así, en Sevilla, por ejemplo, con el 43% de la población activa en los servicios, la productividad, y las ganancias, en el sector son mucho más bajas que las medias nacionales. Mientras tanto, en otras provincias como Madrid o Baleares, a una población activa en los servicios muy elevada —58% y 53% respectivamente— corresponden productividades y salarios y beneficios mucho más elevados que los promedios nacionales. Resumiendo: si los salarios medios en los servicios son muy desiguales en el país, se debe básicamente al hecho de que ciertos subsectores de los servicios tienen unos beneficios sumamemte altos, y se distribuyen por el país de acuerdo con una normativa distinta a la importancia relativa del sector en cada provincia. Habría pues que descubrir cuál es esa normativa.

Entre los otros coeficientes del cuadro 2, interesa destacar el hecho que tanto el porcentaje de población activa en el sector primario, como el correspondiente al sector secundario, son variables altamente correlacionadas con la renta per cápita, respectivamente de forma negativa y positiva. En el primer caso se debe a que se trata de un sector bastante uniforme en sí mismo cuya productividad y beneficios son muy bajos; así, las provincias más ruralizadas se corresponden bastante bien con las de renta más baja, y viceversa. Es decir, que el sector agrario recibe por su trabajo unos beneficios muy bajos, y que éste suele estar distribuido, en términos relativos, por la superficie nacional de una manera muy poco uniforme, y de forma casi tan irregular como la inversa de la distribución de la renta per cápita. También es

Cuadro 2

COEFICIENTES DE CORRELACION ENTRE LA RENTA PER CAPITA PROVINCIAL (MEDIA) Y LAS TASAS PROVINCIALES DE ACTIVIDAD EN CADA SECTOR

Variables correlacionadas	Coef. de correlac. "r"	Coef. de determinac.
Renta p. c. — Poblac. act. sector I (°/o)	-0,83	0,69
Renta p. c. — " " " II (°/o)	0,86	0,74
Renta p. c. — " " " III (°/o)	0,41	0,17
°/o Pobl. act. sect. — °/o Pobl. act. sect. II (1)	-0,80	0,64

(1) Excluida la actividad de la construcción.
Fuente: Base en: Banco de Bilbao: "La renta... 1973"

interesante resaltar aquí, que la distribución de la población activa en este sector está fuertemente reñida con la distribución de la población activa en la industria; es decir, donde un sector predomina, el otro tiende fuertemente a tener poca importancia. Finalmente, es interesante ver cómo la población activa en el sector secundario varía, en su distribución proporcional, a través de la superficie nacional, de tal forma que "casi" se corresponde con la variación de la variable renta per cápita a través de la misma superficie. Es decir, que el hecho "industrialización" implica, o conlleva una serie de características que en conjunto nos "explican" nada menos que el 74% de la variación espacial de la renta per cápita.

Así las cosas, hemos obtenido un nuevo coeficiente de correlación que puede aclarar bastante la situación. Se trata de las variables "densidad de la renta" (renta provincial total / superficie de la provincia) y "densidad de trabajo" (número total de empleos de la provincia / superficie de la provincia), que resultan estar correlacionadas positivamente con un coeficiente "r" de 0,79. Es decir, que una variable "explica" el 62% de la variación espacial de la otra, o, que ambas variables se "explican mutuamente" en el mismo porcentaje. Así, el factor "concentración superficial de la actividad" está muy relacionado con el factor "concentración superficial de la renta".

Resumiendo. Después de analizar el detalle provincial de la variación de todas las variables mencionadas anteriormente (21), podemos concluir de la forma siguiente:

1.º El factor "agrarización", o "ruralización" tiene una incidencia muy negativa en la distribución de la renta. Este tiende a predominar en las áreas del país con los niveles de renta más bajos.

2.º El factor "industrialización" incide muy positivamente en la distribución de la renta. Tiende a concentrarse precisamente en las áreas del país con mayores niveles de renta, y tiende a ser muy poco importante en las áreas más pobres.

3.º La actividad de los Servicios tiene una incidencia positiva, aunque no muy grande, en la distribución de la renta. Su productividad tiende a ser muy alta precisamente en las áreas más ricas, y ello se debe básicamente a que esas áreas concentran la actividad de servicios de más calidad, es decir, lo que se viene conociendo como actividad cuaternaria.

4.º La "densidad superficial del esfuerzo laboral" está muy relacionada, y de forma positiva, con la distribución de la renta per cápita a través de la superficie nacional.

En conjunto, cualquier tipo de actividad, incluida la agraria, tiende a ser relativamente más productiva en las áreas más ricas y de mayor densidad laboral. En esas áreas se concentran, además, las actividades de mayor "calidad" en cualquiera de sus gamas, al menos desde el punto de vista de la rentabilidad y los beneficios. En resúmen, todo esto pone claramente de manifiesto lo que ya se ha estudiado detalladamente para tantos otros paises y regiones: las *economías de concentración* rompen el equilibrio general de los marcos geográficos en que se sitúan; en el interior de estas áreas, la riqueza parece perpetuarse y multiplicarse con suma facilidad, mientras que las áreas marginales, sometidas al mecanismo de intercambio desigual y privadas de los beneficios de todo tipo que se disfrutan en las zonas de economía concentrada (comunicaciones, cultura, relaciones personales, posibilidades, etc.), "gozan" de una renta relativamente baja y tienen escasas posibilidades de "competir" con las áreas "concentradas" para mejorar sustancialmente su relativamente mala posición económica.

(1) Este artículo formaba inicialmente parte de otro más extenso en que se combinaban las técnicas de correlación y regresión lineales y simples para estudiar la distribución espacial de la renta en España. Por razones de falta de espacio tuvo que ser descompuesto en dos. El segundo, "Sobre el uso de la regresión lineal simple en Geografía. Aplicación al estudio de la distribución espacial de la renta per cápita en España" puede encontrarse en "Paralelo 37. Revista de Geografía", núm 1, pp. 83-102, Almería, 1977. (Por sus numerosas erratas de imprenta consulte Fe).

(2) Variable es una magnitud que varía, tanto en el tiempo como en el espacio. Ejemplos de variables podrían ser la temperatura, altura sobre el nivel del mar, las inversiones en la agricultura a través de la superficie de un país, o la evolución de la emigración de una provincia en una serie determinada de años.

(3) Una función es una expresión matemática que mide la relación existente entre varias variables. Si decimos que en una región las lluvias dependen de la altitud, estamos expresando una función en que la primera variable "depende" de la segunda. Esto podría estar expresado mediante una fórmula matemática del tipo $Y_i = a \pm b \cdot X_i + e$ dónde Y_i representa las precipitaciones de cada uno de los puntos del área mencionada; "a" y "b" representan valores constantes, específicos para ésta función; X_i representa la altitud de cada uno de los puntos de observación, y "e" representa un error o un valor incontrolado, no especificado en la función.
Una función exponencial se representa por una ecuación general del tipo $Y_i = a \pm b \cdot X_i^n$ En ella, los valores que en cada caso va tomando la variable Y dependen de la potencia enésima de los valores correspondientes de la variable X. Así, los valores de Y no varían de una forma uniforme y constante, sino progresivamente creciente o decreciente. Una función de este tipo en que Y crece de una forma progresivamente creciente, se representaría en un gráfico por una línea del tipo de la que aparece en la figura 1— c.

(4) El concepto de covarianza descansa, a su vez, en el de *varianza*. La varianza es una medida estadística de dispersión que describe matemáticamente la "variación experimentada por una variable en sí misma". Mide el grado de dispersión de una serie de valores en torno a su media aritmética, y se define como la media del cuadrado de las desviaciones sobre la media. Su fórmula es:

$$\text{Varianza} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad \text{(F-6)}$$

(5) La desviación estándar, o desviación típica, es la raiz cuadrada de la varianza. Su fórmula es:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \qquad \text{(F-7)}$$

Para cálculos rápidos, con calculadora manual simple, puede ser sustituida por:

$$\sigma_x = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} \qquad \text{(F-8)}$$

La desviación estándar es una medida relativa de la dispersión que depende de la magnitud de los datos en que aparece medida la variable. Si esta es grande, también lo serán las desviaciones de los valores observados con respecto a su media. Una medida más "inteligible" de la desviación típica sería el porcentaje de la misma con respecto a la media. La desviación estándar es una medida abstracta muy utilizada en Estadística pues, siendo relativa, sirve para establecer comparaciones entre variables medidas en magnitudes muy diferentes.

El "contenido" profundo de esta medida es el siguiente: si los elementos de la variable proceden de una distribución "normal"(ver nota (8)), el 68,3% de las ocurrencias u observaciones de la variable, deberán tener una desviación sobre la media inferior al valor de una unidad de desviación estándar (estarían comprendidos entre $+1{,}0\sigma$ y $-1{,}0\sigma$). El 95,45% de las observaciones de la variable tendrían una desviación sobre la media comprendida entre ± 2 unidades de desviación estándar. El 99,7% de las observaciones tendrían una desviación inferior a ± 3 unidades de desviación estándar. El 99,99% de las observaciones tendrían una desviación sobre la media inferior a $\pm 4{,}0\sigma$.
Ejemplo de obtención.— Supongamos que una finca, en 8 años consecutivos ("n"=8), ha producido, respectivamente, 6,6,7,8,4,5,6 y 9 toneladas de naranja, y queremos hallar la desviación típica de la producción de esa serie de años. Para poder aplicar la fórmula F—7 necesitamos conocer: 1.º) "n" o número de datos computados = 8.; 2.º) "$\bar{x}$", o media aritmética de la producción de todos los años = (6+6+7+8+4+5+6+9) / 8=6,35.; 3.º) $\sum_{i=1}^{n} (x_i - \bar{x})^2$ es decir, la suma del cuadrado de las diferencias entre la producción de cada año y la media de los 8 años = $(6-6{,}37)^2 + (7-6{,}37)^2 + (8-6{,}37)^2 + (4-6{,}37)^2 + (5-6{,}37)^2 + (6-6{,}37)^2 + (9-6{,}37)^2 = (-0{,}37)^2 + (-0{,}63)^2 + (1{,}63)^2 + (-2{,}37)^2 + (-1{,}37)^2 + (-0{,}37)^2 + (2{,}63)^2 = 0{,}14 + 0{,}14 + 0{,}39 + 2{,}65 + 5{,}61 + 1{,}88 + 0{,}14 + 6{,}92 = 17{,}87$.
Ahora ya estamos en condiciones de sustituir nuestros datos en la fórmula F—7:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} = \sqrt{\frac{17{,}87}{8}} = \sqrt{2{,}23} = 1{,}49$$

(6) Hay tres tipos de escala de medida de datos: a) *Nominal* o cualitativa; los datos de esta escala no son cuantificables; están divididos en categorías que se excluyen mutuamente y tienen la misma importancia. Ejemplo de este tipo de datos podrían ser los conceptos Navarra, Zaragoza, Valladolid y Almería. b) *Ordinal* en esta escala de medida, los datos aparecen según un único criterio: el lugar que éstos ocupan en una lista que previamente ha sido ordenada de acuerdo con algún criterio. Así, en el ejemplo anterior, las cuatro provincias podrían ser valoradas en una escala ordenada, según su renta per cápita con las puntuaciones 1,2,3 y 4, respectivamente. y c) *Con intervalo*; en esta escala, cada valor viene expresado por una magnitud, medida en una escala contínua, de forma que la distancia entre cada uno de ellos tenga un valor conocido; así, la medida de una superficie en Has.: si una parcela tiene 0,8 Has., sabemos que equivale exactamente al doble de otra de 0,4 Has., o a 0,8 veces otra de 1,0 Has. Los datos del tercer tipo pueden ser cambiados de magnitud de medida sin perder sus cualidades. Así, 0,8 Has. equivalen a 0,008 km^2 0,4 Has. equivalen a 0,004 km^2., y 0,008 km^2. siguen siendo el doble de 0,004 km^2 Por estas propiedades, los datos medidos en esta última escala son los más útiles en Estadística. Sólo las variables medidas en ella pueden tener una media o una desviación típica.

(7) Cuando la relación es no lineal y está poco marcada como tal (se aproxima a una relación lineal), también se puede usar el coeficiente de correlación lineal; éste mide la relación entre ambas variables desde el punto de vista de una hipotética relación lineal y, por ello, en un caso de estos daría un coeficiente más pequeño, aunque, en todo caso, muy próximo al coeficiente de correlación no lineal correspondiente, que sería el más adecuado. Cuando la relación es patentemente no lineal, podemos convertirla en otra de tipo lineal mediante un sistema llamado *transformación*. Hay muchas formas de realizar las transformaciones, que dependen de las características específicas de la relación en cuestión. Las dos más corrientes en Geografía transforman en lineales funciones en que la variable dependiente varía de una forma progresivamente mayor o menor por cada unidad de cambio de la variable independiente, y viceversa; estos casos se pueden representar por medio de funciones exponenciales o logarítmicas. Si, por ejemplo, transformando todos los valores de una de las variables en sus logaritmos correspondientes obtenemos una relación lineal de lo que había sido una relación no lineal, y el coeficiente de correlación lineal obtenido es de, supongamos, 0,8, podríamos decir que la variación de una de ellas (la no transformada, por ejemplo) está "explicada" en un 64% $(0{,}8^2 \cdot 100)$ por la variación de la otra, y viceversa. (El mecanismo de transformación de las principales funciones, así como su problemática, puede verse en cualquier texto de Estadística general).

(8) Los elementos de una variable, o las distintas observaciones de una muestra, se distribuyen "normalmente" cuando sus valores aparecen dispuestos de forma simétrica en torno a la media y su número disminuye progresivamente a medida que nos alejamos de la misma. Si los representamos en un histograma de frecuencias en cuyo eje horizontal aparecen las "clases modales" con sus "límites de clase" correspondiente (por ejemplo, grupos de edades de una población), y en el vertical aparece la "frecuencia", o "número de ocurrencias de cada clase" (por ejemplo, el número total de personas dentro de cada "clase modal"), deberíamos de obtener una curva en forma de campana.
En el caso de una población humana, como la de nuestro ejemplo, no obtendríamos una curva de frecuencias distribuida normalmente porque las clases modales de la izquierda (grupos de jóvenes) tendrían unas frecuencias muy superiores (mayor número de individuos) a las correspondientes a las clases de la derecha (grupos de adultos y viejos). Un ejemplo de distribución normal sería el de una población de adultos dispuestos en grupos (clases modales) según su estatura; en este caso, los individuos pequeños y los altos aparecerían en proporciones pequeñas y de la misma magnitud aproximadamente; los individuos de estaturas medias aparecerían en el centro del histograma y con frecuencias muy grandes.
Un gran número de técnicas estadísticas de interés geográfico (técnicas paramétricas) necesitan basarse en datos procedentes de poblaciones normalmente distribuidas. En Geografía, como en las ciencias sociales, se presentan corrientemente muchas distribuciones "casi" normales, generalmente asimétricas, como podría ser la distribución de los individuos de una comunidad rural andaluza, según sus niveles de renta; en este caso, la mayor parte de las ocurrencias u observaciones tenderían a situarse en la parte izquierda de la curva (grupos de individuos con ingresos más bajos), mientras que las clases modales de la derecha tendrían frecuencias muy pequeñas (pocos individuos con elevados ingresos, que a veces pueden ser realmente elevados).
En las ciencias sociales se suelen aceptar como normales muchas distribuciones que no son exageradamente asimétricas, y, también, otras distribuciones asimétricas cuando se supone que la asimetría se debe a una información mal tomada o insuficiente.

Hay muchos procedimientos para saber si una distribución determinada es o no "normal". Uno de los más simples y utilizados se conoce como el del *"grado de perfección del ajuste"* (*"Goodness-of-fit"*). Este se basa en la prueba, o test, estadística χ^2 ("Chi cuadrado") para dos muestras en que se acepta el nivel crítico o de rechazo de 0,1. Se basa en la comparación de la distribución en cuestión (observada) con lo que sería una distribución normal perfecta (esperada). La hipótesis nula (H_o) se define así: ambas distribuciones son idénticas; se rechaza esta hipótesis, y se acepta la hipótesis alternativa (H_1) (es decir, la que postula que ambas distribuciones son distintas), cuando existe una probabilidad estadística de que H_o sea cierta en menos del 10% de los casos. En caso contrario, se acepta H_o.
Esta prueba se puede hacer de muchas formas. La más

simple consiste en lo siguiente: 1.º) se estandarizan los valores de la distribución que estamos probando mediante la aplicación de la fórmula

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \qquad (F\text{-}9)$$

(dónde z_i representa cada uno de los "n" valores de la variable X, estandarizados. x_i representa a cada uno de los valores de la variable X. σ_x es la desviación estándar de la misma variable). 2.º) se dividen los valores estandarizados en cuatro grupos, o clases modales — podrían dividirse en más grupos, por supuesto — que incluyen respectivamente los valores comprendidos entre los límites $-\infty$; $-0{,}67$; $0{,}0$; $+0{,}67$ y $+\infty$ Puesto que sabemos que en una distribución normal estos límites dividen a la población en cuatro partes iguales, podemos dividir los datos de nuestra distribución, ya estandarizados, en estos cuatro grupos, y comparar el número de valores "observados" de cada grupo con el número de valores "esperados" (que habría en cada grupo si la distribución fuese normal). Esta comparación, como hemos dicho, se hace mediante la prueba χ^2 Su fórmula es:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \qquad (F\text{-}10)$$

(dónde o_i representa el número de valores "observados" dentro de cada una de las "i" clases (cuatro clases). e_i es el número de valores "esperados" en las clases correspondientes (el 25% del total en nuestro caso).
Una vez efectuada la operación obtenemos un valor que si resulta ser inferior a 2,71 nos indica que hemos de aceptar por normal la distribución en cuestión, porque hay más del 90% de probabilidad que se cumpla la hipótesis nula. (ver Davis, J. (1973), pp. 114—123). Existen numerosos procedimientos para transformar los datos de una distribución no "normal" en otros distribuidos de forma aproximadamente normal, y, por tanto, susceptibles de utilizar en técnicas paramétricas (ver Gregory, S. (1975), pp. 52 y sigs.

(9) Hammond, R. & McCullagh, P.S. (1975), pp. 196.

(10) Dalton *et. al.* (1976), pp. 34.

(11) En todo caso, si la correlación no fuese perfecta, se aproximaría mucho a ella; ello se debería a que no todas las familias tienen igual número de componentes ni tampoco igual número de trabajadores.

(12) Ver Taylor, P.J. (1977), pp. 221. Este fenómeno es consecuencia de que cuando aumentamos la extensión de la unidad superficial, eliminamos una parte de la variación de la variable, sobre todo, la que se refiere a interferencias locales de otros factores, mientras que la variación general, que investigamos, permanece casi entera debido al fenómeno, generalmente presente en Geografía, de la autocorrelación espacial (tendencia a la uniformidad en las áreas contiguas).

(13) robinson, A.H. (1956). Este tipo de solución ha demostrado ser válido sólo para cierto tipo de situaciones geográficas. Su mayor problema surge del criterio seguido para efectuar el pesado, sujeto generalmente al sujetivismo del autor.

(14) Taylor, P.J. (1977), pp. 223. Aquí queremos acentuar que éste es uno de los problemas más fuertes en que se basan los detractores de la Geografía Cuantitativa. Siguiendo estrictamente la teoría estadística, no se podrían ajustar las técnicas a la inmensa mayor parte de las situaciones que se presentan en Geografía.

(15) Si tomásemos la unidad superficial de la región, probablemente los coeficientes de correlación serían algo más altos. Para la elección de la unidad provincial, hemos tenido en cuenta que algunas de las regiones no son demasiado uniformes en sí mismas, a la vez que, las provincias suelen ofrecer mayor coherencia interna y, además, pueden darnos una visión mucho más rica en matices.

(16) Aquí queremos recordar que, con 50 provincias, el coeficiente de correlación gradual puede estar bastante deformado ya que, como vimos, sólo resulta dar buenos resultados para casos en que las variables tienen entre 10 y 30 valores.

(17) La elección de una u otra unidad de superficie, recuerda Taylor, sería equiparable teóricamente a la elección de diferentes tipos de "muestras", y, por tanto, estaría sometida a un cierto error. De esta forma, nuestros resultados no serían absolutamente exactos, sino sólo, "muy aproximados" a los que obtendríamos si trabajásemos con "todos" los elementos de la población estadística a los que, se supone, representa la muestra. En nuestro ejemplo, la elección de la unidad superficial de la provincia nos expone a un cierto error estadístico; pensemos que, en España, no pueden tener la misma significación provincias como Guipúzcoa y Badajoz: la segunda es unas 10 veces más extensa que la primera, y tiene una densidad de población diez veces inferior.

(18) Gregory, S. (1975), pp. 198.

(19) Si hubiésemos obtenido un coeficiente de correlación del orden, por ejemplo, de 0,8, para un caso con 48 grados

de libertad, obtendríamos en el mencionado gráfico un punto "b" que, al estar situado por encima del nivel de confianza del 0,1%, nos indicaría que hay menos del 0,1% de probabilidad de haber obtenido el coeficiente por "error", y, por tanto, habría que rechazar la hipótesis nula para aceptar, con plena confianza, la hipótesis alternativa, esto es, que la mencionada correlación es auténtica.

(20) Gregory, S. (1975), pp. 192.

(21) El estudio más detallado de éstas, y otras variables, relacionadas con la distribución de la renta per cápita provincial española, puede contemplarse con mayor riqueza de detalle en Compán, D. (1977).

BIBLIOGRAFIA

CLIFF, A.D., HAGGETT, P., ORD, J.K., BASSETT. K.A. & DAVIES, R.B. (1975): "Elements of Spatial Structure. A Quantitative approach", Cambridge Univ. Press, 258 pp.

COLE, J.P. (1975): "Una introducción al estudio de métodos cuantitativos aplicados en Geografía", Univ. Autónoma de Mexico, 94 pp.

COLE, J.P. & KING, C.A.M. (1970): "Quantitative Geography. Techniques and Theories in Geography", John Wiley & Sons Ltd., London, 692 pp.

COMPAN VAZQUEZ, D.(1977): "Sobre el uso del análisis De la regresión lineal simple en Geografía. Aplicación al estudio de la distribución espacial de la renta de España", en *Paralelo 37. Revista de estudios geográficos*. Almería, núm. 1; 83-102.

DALTON, R., GARLICK, J. MINSHULL, R. & ROBINSON, A. (1976): "Correlation Techniques in Geography", George Philip & Son Ltd., London, 60 pp.

DAVIS, J.C. (1973): "Statistics and Data Analysis in Geology", John Wiley & Sons Inc., N. York, 550 pp.

EBDON, D. (1977): "Statistics in Geography", Basil Blackwell, Oxford, 195 pp.

EZEQUIEL, M. (1949): "Methods of Correlation Analysis", John Wiley, 2.ª ed., 531 pp.

GREGORY, S. (1975): "Statistical Methods and the Geographer", Longman, 3.ª ed., 270 pp.

GROUPE CHADULE (1974): "Initiation aux méthodes statistiques en Géographie", Masson et Cie., Paris, 192 pp.

HAMMOND, R. & McCullagh, P.S. (1975): "Quantitative Techniques in Geography", Oxford Univ. Press, London, 318 pp.

HOEL, P.G. (1976): "Introducción a la Estadística matemática", Ariel, Barcelona, 431 pp.

KING, L. (1969): "Statistical Analysis in Geography", Prentice-Hall Inc., Eglewood Cliffs, N. Jersey, 288 pp.

LOPEZ CACHERO, M. (1976): "Fundamentos y métodos de Estadística", Ediciones Pirámide, Madrid, 574 pp.

MAXWELL, A.E. (1977): "Multivariate Analysis in Behavioural Research", Chapman and Hall, London, 164 pp.

McCULLAGH, P. (1974): "Multivariete Analysis in Beha- Oxford Univ. Press, 120 pp.

MEYER, D.R. (1971): "Factor Analysis versus Correlation Analysis: are substantive interpretations congruent?", in *Economic Geography*, 47, 336—343.

POOLE, M.A. & O'FARREL, P.N. (1971): "The assumptions of the linear regression model" in *Transactions*, 52, 145—159.

ROBINSON, A.H. (1956): "The Necessity of Weighting in Correlation of Areal Data", in *Annals Assoc. Am. Geographers*, 46, 233—236.

SMITH, D.M. (1971): "Industrial Location. An Economic Geographical Analysis", John Wiley & Sons Inc., N. York, 553 pp.

SPIEGEL, M.R. (1975): "Estadística. Teoría y 875 problemas resueltos", Ediciones de la Colina, Madrid, 357 pp.

TAYLOR, P.J. (1977): "Quantitative Methods in Geography. An Introduction to Spatial Analysis", Houghton Mifflin Co., Boston, 386 pp.

TILL, R. (1974): "Statistical Methods for the Earth Scientist. An Introduction", MacMillan, Norwich, 154 pp.