

Unidade 10: Regresión

- 1 Variables bidimensionais**
 - 1.1 Táboas de frecuencias.**
 - 1.2 Nube de dispersión.**
 - 1.3 Covarianza.**
- 2 Relación entre as variables**
 - 2.1 Nubes de puntos**
 - 2.2 Regresión. Regresión lineal**
 - 2.3 Correlación. Coeficiente de correlación lineal.**

Introducción

Funcións

Lembra que definimos unha función como unha relación entre magnitudes numéricas que verifica certas propiedades.

Podemos describir unha función empregando:

- Expresións verbais.
- Táboas de valores.
- Gráficas.
- Fórmulas.

Exercicio 10.1: Observando os recibos eléctricos vemos que os únicos datos que varían dun a outro son o consumo (en Kw·h) e o importe (en €). Podemos supoñer que o importe de cada recibo calcúlase a partir do consumo pero ¿de que xeito?. Intentemos descubri-lo a partir dos datos dalgúns recibos¹:

Consumo (Kw·h)	359	423	453	388
Importe (€)	47	53,4	56,4	49,9

- a) Representa graficamente eses valores.
- b) ¿Que tipo de gráfica se obtén?.
- c) Calcula a fórmula da función correspondente a esa gráfica.
- d) Comproba que todos os valores da táboa axústanse a esa fórmula.
- e) ¿Cal será o importe para un consumo de 500 Kw·h?
- f) ¿Que significado “real” teñen os coeficientes que aparecen nesa fórmula?

¹ Os datos son reais. Só están lixeiramente modificados para evitar problemas co redondeo e facilitar os cálculos.

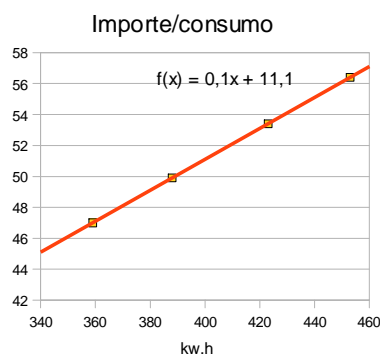
Unha situación menos excepcional

No exercicio anterior describimos unha función, a relación entre o consumo e o importe dos recibos eléctricos, mediante unha táboa.

Podemos comprobar que os puntos da táboa corresponden exactamente aos valores da fórmula que describe a función, tal como se aprecia na gráfica, pero esa non é a situación máis habitual.

Na maioría das ocasións, ou ben os datos conteñen algún erro experimental, ou ben a dependencia entre as variables é só “aproximada”, ou suceden as dúas cousas.

Este tipo de relación entre magnitudes non podemos estudiala cun enfoque analítico. Debemos buscar a función que “mellor se adapte” aos puntos. Necesitamos un novo enfoque.



Exercicio 10.2: Co obxecto de planificar axeitadamente o curso 2002, quérese investigar cantos alumnos estarán matriculados na Universidade de Santiago de Compostela nese ano.

O número de matriculados en cursos anteriores aparece na seguinte táboa:

Curso	1997	1998	1999	2000
Matriculados (en miles) ²	42,414	42,926	42,121	41,654

- Representar graficamente eses valores.
- ¿Que tipo de gráfica se obtén?
- Coa axuda dunha regra, debuxa a recta que mellor se adapte a eses puntos e calcula a fórmula da función correspondente a esa gráfica.
- ¿Cantos alumnos se matricularán no ano 2002?

² Fonte: anuario El Pais

Regresión

Variables bidimensionais

Frecuenteente estudaremos dúas variables diferentes nunha mesma poboación (idade e preferencias políticas, estatura e peso, notas en dúas materias, consumo alcohólico e número de accidentes, etc).

Chamámoslle variable **estatística bidimensional** a distribución conxunta das dúas variables.

Representando graficamente (cada punto por separado) os valores da distribución bidimensional obtemos a **nube de puntos** da distribución ou **diagrama de dispersión**.

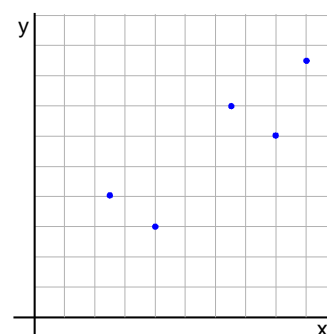
Exemplo: As cualificacións en dous exames de Matemáticas de 5 alumnos foron:

1º exame	2.5	4	6'5	8	9
2º exame	4	3	7	6	8'5

Representa a nube de puntos e calcula as medias e desviacións típicas desas variables (fíxate que a frecuencia de cada resultado é 1).

Solución: Poñemos os datos nunha táboa para facilita-la súa manipulación:

$$\bar{x} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{30}{5} = 6 \quad \bar{y} = \frac{\sum_{i=1}^5 y_i}{5} = \frac{28'5}{5} = 5'7$$
$$s_x^2 = \frac{\sum_{i=1}^5 x_i^2}{5} - \bar{x}^2 = \frac{209'5}{5} - 6^2 = 5'9 \rightarrow s_x = \sqrt{5'9} = 2'43$$
$$s_y^2 = \frac{\sum_{i=1}^5 y_i^2}{5} - \bar{y}^2 = \frac{182'25}{5} - 5'7^2 = 3'96 \rightarrow s_y = \sqrt{3'96} = 1'99$$



x_i	y_i	x_i^2	y_i^2
2.5	4	6'25	16
4	3	16	9
6'5	7	42'25	49
8	6	64	36
9	8'5	81	72'25
30	28'5	209'5	182'25

Regresión e correlación

Nunha variable estatística cualitativa bidimensional tratamos de coñecer se existe algunha relación entre os valores dos dous caracteres (notas en dous exames; estatura e peso; consumo alcohólico e número de accidentes, etc.).

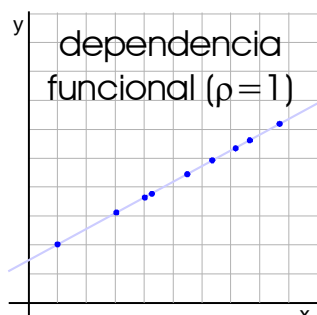
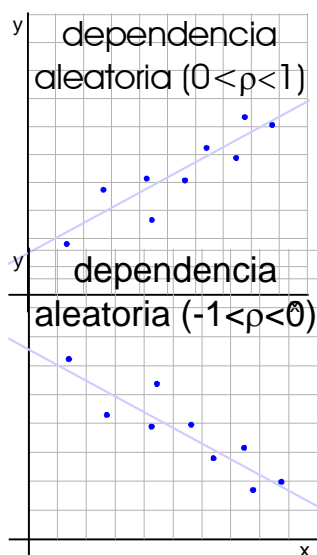
Para estudar a relación entre as variables necesitamos “medir” a variación conxunta das mesmas. Para facelo ideouse a **covarianza**:

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

Fíxate que se un individuo ten os datos altos en relación ás medias, aportará para a covarianza ($+ \cdot + = +$) e tamén se os seus datos son baixos ($- \cdot - = +$), pero se os datos non se corresponden entre si perxudicará á covarianza ($+ \cdot - = -$ e $- \cdot + = -$).

De xeito semellante á como faciamos na varianza, dispoñemos dunha segunda fórmula para o cálculo da covarianza:

$$S_{xy} = \frac{\sum_{i=1}^n x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$



Correlación lineal

É o estudio do grao de dependencia lineal entre os caracteres dunha variable bidimensional.

Para medir ese grao de dependencia utilízase o coeficiente de **correlación lineal**: $\rho = \frac{S_{xy}}{S_x \cdot S_y}$

O coeficiente de correlación lineal sempre comprendido entre -1 e 1 .

Poden darse diferentes graos de dependencia lineal entre os caracteres dunha variable bidimensional.

1. **Dependencia aleatoria** ou **probabilística**: A nube de puntos aproxímase a unha recta. O coeficiente de correlación ten un valor entre 0 e 1 (ou entre -1 e 0).

Se é próximo a 1 indica que hai moita dependencia, directa (se sobe un carácter sobe tamén o outro, parello a el).

Se é próximo a -1 indica que hai moita dependencia inversa (se sobe un caracter o outro baixa).

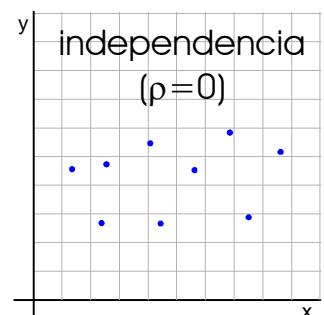
Canto máis próximo a 0 menor grao de dependencia.

- 2. Dependencia funcional:** É un caso extremo cando hai unha dependencia total entre as variables.

A nube de puntos está sobre unha recta. O coeficiente de correlación é 1 (recta crecente) ou -1 (decrecente).

- 3. Independencia:** Cando non hai relación lineal entre as variables (pode haber unha relación doutro tipo).

A nube de puntos non se asemella a unha recta. O coeficiente de correlación é 0.



Exemplo: O peso e a estatura dun grupo de persoas son:

Peso (kg)	55	60	64	68	70	70	72	81
Estatuta (cm)	162	166	166	175	168	174	170	173

Representar a nube de puntos e estudar o grao de dependencia lineal entre as variables peso e estatura.

Solución: O grao de dependencia ven dado polo coeficiente de correlación lineal, polo que debemos calcular as medias, desviacións típicas e covarianza das variables:

$$\bar{x} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{540}{8} = 67.5 \text{ kg} \quad s_x = \sqrt{\frac{\sum_{i=1}^8 x_i^2}{8} - \bar{x}^2} = \sqrt{\frac{36890}{8} - 67.5^2} = 7.4 \text{ kg}$$

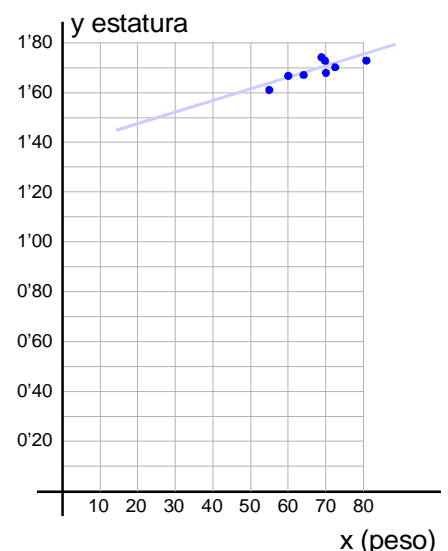
$$\bar{y} = \frac{\sum_{i=1}^8 y_i}{8} = \frac{1354}{8} = 169.25 \text{ m} \quad s_y = \sqrt{\frac{\sum_{i=1}^8 y_i^2}{8} - \bar{y}^2} = \sqrt{\frac{229310}{8} - 169.25^2} = 4.26 \text{ m}$$

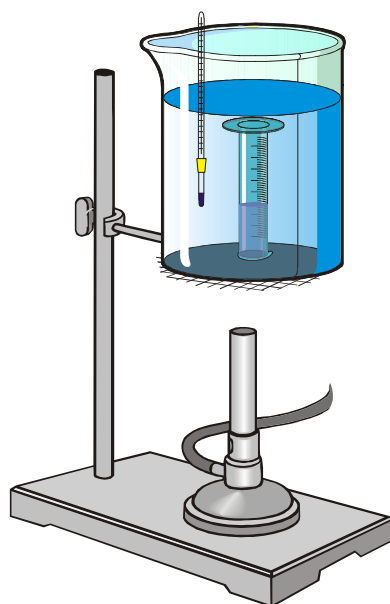
$$s_{xy} = \frac{\sum_{i=1}^8 x_i \cdot y_i}{8} - \bar{x} \cdot \bar{y} = \frac{91587}{8} - 67.5 \cdot 169.25 = 24$$

O coeficiente de correlación lineal será: $\rho = \frac{24}{7.4 \cdot 4.26} = 0.76$

0.76 é un valor entre 0 e 1, o que indica un certo grao de relación lineal positiva (crecente) entre as variables. Hai tendencia a que as persoas de maior peso sexan máis altas e as de menor peso sexan máis baixas pero non é radicalmente así e son bastante frecuentes os casos en que non ocorre.

0.76 indica unha relación lineal importante, pero non moita (non é aínda determinante, tería que aproximarse máis a 1).





Exercicio 10.3: Medimos o volume dun gas (aire) a diferentes temperaturas, mantendo fixa a presión, como se indica na figura. Os datos obtidos foron:

Temperatura (°C)	15'4	20'1	24	28	32'6	36
Volume (cm ³)	15'5	15'7	16	16'4	16'9	17'2

- Representar eses valores.
- Determina a ecuación da recta que cres que se axusta mellor a eses datos.
- Atopa, utilizando a túa recta, a temperatura que corresponde a un volume 0. ¿Cal é o significado físico desta temperatura?

Regresión lineal

Se unha nube de puntos se poidese aproximar por medio dunha recta, esta facilitaríanos enormemente o seu estudio (aínda que perderíamos algunha exactitude).

Chámase regresión lineal á búsqueda da fórmula lineal ($y=ax+b$) que mellor se axuste a nube de puntos dunha distribución bidimensional.

Pode haber varias rectas “parecidas” que nos resulten de utilidade: segundo cal sexa o criterio que utilizemos seleccionaremos unha ou outra.

O criterio máis estendido é o chamado *de mínimos cadrados* que utiliza as diferenzas de ordenadas (ou abscisas) que hai entre os puntos da nube e os da recta (as diferenzas entre os valores reais e os valores que daría a recta).

Hai dúas rectas de mínimos cadrados (a correspondente as diferencias de ordenadas e a das diferencias de abscisas):

$$\text{Recta de Y sobre X: } y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

$$\text{Recta de X sobre Y: } x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

Exemplo: Imos calcular a recta de regresión correspondente ao exercicio anterior:



$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{156'1}{6} = 26'016 \text{ }^{\circ}\text{C} \quad s_x = \sqrt{\frac{\sum_{i=1}^6 x_i^2}{6} - \bar{x}^2} = \sqrt{\frac{4359'93}{6} - 26'016^2} = 7'056 \text{ }^{\circ}\text{C}$$

$$\bar{y} = \frac{\sum_{i=1}^6 y_i}{6} = \frac{97'7}{6} = 16'283 \text{ cm}^3 \quad s_y = \sqrt{\frac{\sum_{i=1}^6 y_i^2}{6} - \bar{y}^2} = \sqrt{\frac{1593'15}{6} - 16'283^2} = 0'615 \text{ cm}^3$$

$$s_{xy} = \frac{\sum_{i=1}^6 x_i \cdot y_i}{6} - \bar{x} \cdot \bar{y} = \frac{2567'61}{6} - 26'016 \cdot 16'283 = 4'316$$

O coeficiente de correlación lineal será:

$$\rho = \frac{4'316}{7'056 \cdot 0'615} = 0'995$$

É un valor moi próximo a 1, o que indica un forte grao de relación lineal entre as variables.

Para *estimar* o volume a unha certa temperatura utilizamos a recta de regresión de Y sobre X (volume en función da temperatura):

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \rightarrow y - 16'283 = \frac{4'316}{7'056^2} (x - 26'016) \rightarrow y = 0'087x + 14'028$$

$$\text{Para } x=40: y = 0'087 \cdot 40 + 14'028 = 17'508 \text{ cm}^3$$

Para *estimar* a temperatura a partir dun volume utilizamos a recta de regresión de X sobre Y:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y}) \rightarrow x - 26'016 = \frac{4'316}{0'615^2} (y - 16'283) \rightarrow x = 11'411y - 159'793$$

$$\text{Para } y \approx 0: x = 11'411 \cdot 0 - 159'793 = -159'793 \text{ }^{\circ}\text{C}$$

Esta temperatura é o que se coñece como 0 absoluto: os gases non ocuparían espazo ningún.

Exercicio 10.4: Para estudar a existencia do cambio climático orixinado polas actividades do home, mídese a temperatura media anual na Terra e a concentración de CO₂, obténdose os datos que figuran na táboa adxunta.

Analiza eses datos coa axuda dunha folla de cálculo e explica que conclusións podemos extraer en relación á existencia do cambio climático antropoxénico.

Modelos exponenciais

As funcións do tipo $y=a \cdot b^x$ son as seguintes en importancia, despois das lineais, para a descrición de fenómenos físicos. Podemos utilizar o modelo de regresión lineal para estudar a dependencia exponencial con só tomar logaritmos:

Ano	CO ₂	°C	Ano	CO ₂	°C
1959	315,98	14,04	1982	341,13	14,08
1960	316,91	13,98	1983	342,78	14,33
1961	317,65	14,09	1984	344,42	14,13
1962	318,45	14,05	1985	345,9	14,11
1963	318,99	14,03	1986	347,15	14,18
1964		13,75	1987	348,93	14,34
1965	320,03	13,86	1988	351,48	14,38
1966	321,37	13,93	1989	352,91	14,24
1967	322,18	13,99	1990	354,19	14,48
1968	323,05	13,92	1991	355,59	14,43
1969	324,62	14,01	1992	356,37	14,14
1970	325,68	14,05	1993	357,04	14,18
1971	326,32	13,91	1994	358,88	14,31
1972	327,46	13,95	1995	360,88	14,46
1973	329,68	14,19	1996	362,64	14,38
1974	330,25	13,95	1997	363,76	14,4
1975	331,15	13,97	1998	366,63	14,7
1976	332,15	13,78	1999	368,31	14,45
1977	333,9	14,16	2000	369,48	14,41
1978	335,5	14,07	2001	371,02	14,56
1979	336,85	14,13	2002	373,1	14,69
1980	338,69	14,27	2003	375,61	14,66
1981	339,93	14,39	2004	377,43	14,59

$$y = a \cdot b^x \xrightarrow{\text{tomando logaritmos}} \ln(y) = \ln(a \cdot b^x) \rightarrow \ln(y) = \ln(a) + x \cdot \ln(b)$$

Vemos que é posible expresar a dependencia lineal entre x e $\ln(y)$.

Exemplo: A táboa recolle a poboación española desde 1960 ata 1996.

¿Podemos asegurar que é exponencial?

Solución: Calculamos o logaritmo neperiano de Y :

X	1960	1970	1975	1981	1986	1991	1996
$\ln(Y)$	17'246	17'343	17'399	17'445	17'465	17'476	17'496

O coeficiente de correlación lineal é 0'969, o que indica unha forte relación lineal entre os valores de X e $\ln(Y)$.

Pero observando a gráfica, podemos apreciar que a relación non é tan clara. Os primeiros anos si amosan un crecemento exponencial pero non así os últimos, nos que se observa un certo estancamento cunha leve recuperación no período 1991 a 1996.

A recta de regresión de $\ln(Y)$ sobre x é:

$$y = 0'006885x + 3'77958 \quad [a \text{ y representa o } \ln(Y)]$$

A función exponencial correspondente a Y sobre x será:

$$Y = e^{0'006885x + 3'77958} = e^{3'77958} \cdot (e^{0'006885})^x \rightarrow$$

$$Y = 43'7979 \cdot 1'0069^x$$

(indica un aumento do 0'69% anual)

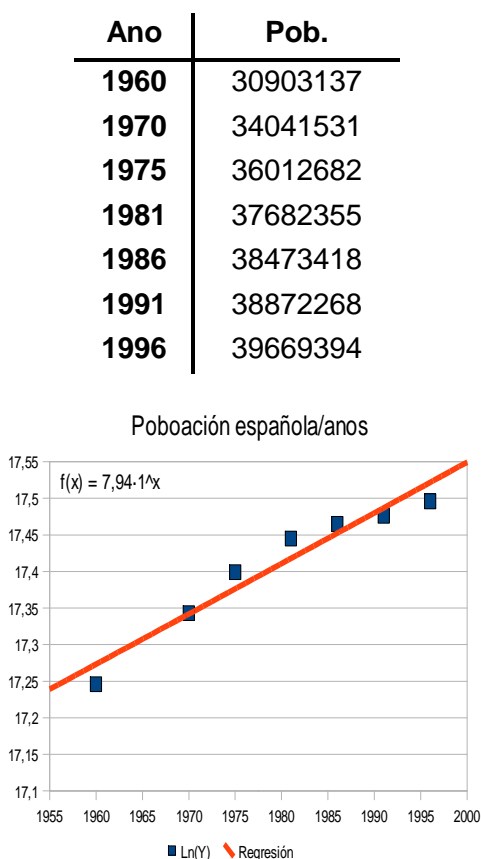
Exercicio 10.6: A xulgar polos datos anteriores, ¿coidas que se pode afirmar un comportamento diferente nos períodos 1960-1975 e 1986-1996?

Exercicio 10.7: A poboación mundial (en millóns de habitantes) nas últimas décadas foi a seguinte:

Año	1950	1960	1965	1970	1976	1983	1990	2000
Población	2502	2987	3288	3610	4044	4700	5300	6054

a) Podemos afirmar que a poboación mundial ten un crecemento exponencial

b) ¿Cuál será la población humana en 2010?



Ampliación

Regresión non paramétrica

Nas situacións anteriores, estudamos o grao de relación lineal entre dúas variables.

O experimentador decide que tipo de función debe empregar e as conclusións dependen desa elección. Este enfoque, fixando o tipo de función de antemán, recibe o nome de regresión paramétrica.

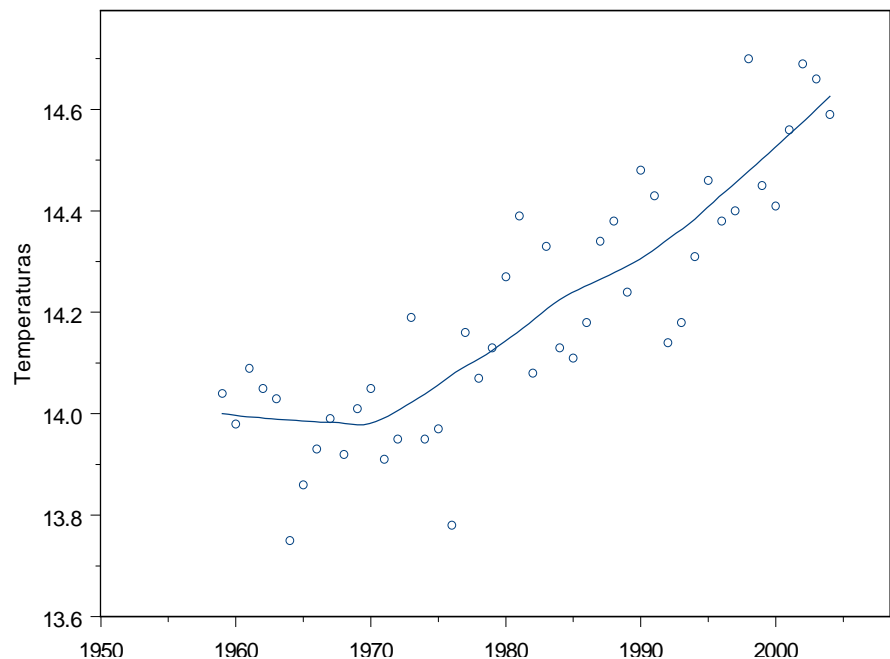
A utilización de ordenadores permite un enfoque diferente, no que o tipo de función non está prefixado. Un enfoque, conceptualmente, máis simple.

Buscamos a relación entre dúas variables, X e Y, das que temos un gran número de datos da súa distribución conxunta, valores que están suxeitos a erros experimentais e a fluctuacións debidas a que a relación entre esas variables non é determinista.

Un xeito de atopar esa relación e, simplemente, asociar a cada valor de X unha certa media dos valores da Y correspondentes a valores cercanos a ese valor de X.

Vemos que sucede ao aplicar este método aos valores da temperatura mundial.

Podemos apreciar que, en 1970 hai unha baixada das temperaturas que non se pode explicar polo aumento do CO₂ pero si por outros factores antrópicos³. A partir dese ano, semella producirse un aumento case lineal da temperatura tal como din os expertos en cambio climático: 0'6° C en 34 anos, uns 0'18° C por década (un valor un pouco superior ao que manexa o IPCC).



³ Hai outros factores antrópicos que si explican esa baixada de temperaturas, como a produción de polvo e aerosois contaminantes que poden impedir a chegada da luz do Sol provocando unha baixada nas temperaturas.

A partir de mediados dos 70 adoptáronse fortes medidas para reducir a produción destes contaminantes que teñen graves efectos na saúde, o que poden desenmascarar o efecto invernadoiro que estaba oculto.