Francesco Lagona

Modelli statistici per l'analisi dell'ambiente

Abstract: I modelli statistici consentono di descrivere la relazione tra variabili ambientali in diversi ambiti, come ad esempio la valutazione del rischio e dell'impatto ambientale, la tutela delle risorse naturali, la prevenzione dell'inquinamento e la previsione di eventi naturali estremi. Il contributo è una breve introduzione ai modelli statistici lineari e offre un esempio di applicazione di tali modelli all'analisi del riscaldamento globale.
Parole chiave: Modelli lineari; riscaldamento globale; statistica ambientale.

Abstract: Statistical models are capable to describe relationships between environmental variables under multiple frameworks, such as environnental risk evaluation, evironmental resources protection, pollution prevention and extreme event forecasting. This contribution is a brief introduction to linear statistical models, illustrated with an application to global warming.
Keywords: Linear models; Global warming; Environmental statistics.

1. *Introduzione*

Il monitoraggio ambientale non si riduce alla mera misurazione di variabili ambientali, ma necessita di metodi che sintetizzino la relazione funzionale tra le variabili osservate. Senza un'adeguata comprensione delle relazioni che intercorrono tra le variabili sono, infatti, impossibili le valutazioni di rischio ambientale, le strategie per la tutela delle risorse naturali, la prevenzione dell'inquinamento e la previsione di eventi estremi.

Consideriamo, per esempio, il problema di individuare la relazione tra il livello della temperatura e la concentrazione di ozono, un noto inquinante atmosferico secondario che si forma in presenza di alte temperature. In laboratorio è possibile generare artificialmente concentrazioni di ozono al variare di differenti valori di temperatura e sotto condizioni sperimentali omogenee, in modo che la variabilità delle concentrazioni sia totalmente imputabile alle variazioni di temperatura. Nel caso di osservazioni sperimentali controllate, è nota la relazione chimica tra temperatura e ozono ed è rappresentabile come una famiglia parametrica di funzioni matematiche del tipo

$$ozono = f(temperatura; \theta)$$

dove θ è un vettore di parametri. L'individuazione della relazione funzionale si riduce dunque alla calibrazione della curva ottenuta mediante il calcolo dei valori dei parametri θ che interpolano in modo ottimale i valori sperimentali osservati. Tale calibrazione viene realizzata attraverso metodi matematici di ottimizzazione.

Più complicata è, invece, l'analisi della formazione dell'ozono nell'aria aperta. Qui intervengono molteplici fattori non osservati (ad esempio la velocità del vento, il livello di radiazione solare, la formazione di inquinanti primari che concorrono alla formazione dell'ozono) e non controllabili che rendono difficile l'individuazione della relazione funzionale tra ozono e temperatura, che infatti non è nota in generale. In situazioni di questo tipo, tipiche del monitoraggio ambientale, un approccio possibile è quello statistico. Secondo questo approccio, le concentrazioni di ozono vengono modellate come determinazioni di una variabile aleatoria Y con distribuzione di densità di probabilità

$$f(y; \theta)$$

nota a meno di parametri che sono legati alla temperatura da una relazione funzionale

$$\theta = g^{-1}(temperatura)$$

Secondo questo approccio, la temperatura influenza solo i parametri di una distribuzione di probabilità attraverso una funzione legame g, che a sua volta genera le concentrazioni di ozono. In altre parole, la conoscenza della relazione funzionale g non ci consente di determinare con certezza i valori di ozono corrispondenti ai valori di temperatura, ma solo la probabilità con cui tali valori compaiono in un intervallo A, calcolata come l'area sotto la densità f che giace sull'intervallo A:

$$P(A) = \int_A f(y; \theta) dy$$

Questo approccio ci consente di separare l'influenza della temperatura sull'ozono dall'influenza dalle molteplici varabili ambientali che non siamo in grado di controllare all'aria aperta, in contesto non sperimentale.

Ogni volta che specifichiamo una distribuzione di probabilità *f* e una funzione legame *g*, specifichiamo un modello statistico. Questa lezione illustra brevemente i modelli statistici lineari che sono i più semplici e più utilizzati in ambito ambientale. Nella teoria dei modelli lineari, la distribuzione *f* è una normale che dipende da un vettore di due parametri

$$\theta = (\mu, \sigma^2)$$

e il legame tra θ i valori di temperatura è di tipo lineare

$$\mu = \beta_0 + \beta_1 \times \text{temperatura} \qquad \sigma^2 = const.$$

2. *La media aritmetica come modello lineare*

I modelli lineari possono essere visti come una generalizzazione del concetto di media aritmetica. Supponiamo di aver a che fare con un campione di osservazioni (ad esempio, misurazioni di ozono) che indichiamo con

$$(y_1, \; \dots y_i \; \dots y_n)$$

e supponiamo che ogni osservazione campionaria sia stata generata in modo indipendente da una normale

$$y_i \sim N(\mu, \sigma^2).$$

Ogni valore y_i può essere interpretato come la somma tra le media μ e un errore ε estratto da una distribuzione normale con media 0 e e varianza σ^2:

$$y_i = \mu + \varepsilon_i \quad (1)$$

Secondo questa interpretazione, tutte le osservazioni condividono lo stesso valore di μ e si distinguono per il diverso valore assunto di volta in volta da ε_i. Per sottolineare questi aspetti, μ prende il nome "componente sistematica o deterministica", mentre ε prende il nome di componente erratica o stocastica. Quando formuliamo un'ipotesi sui dati cercando di distinguere la componente deterministica da quella erratica, diciamo per l'appunto che stiamo specificando un modello statistico. L'equazione

(1) è un modello statistico assai semplice (la Figura 1 ne illustra un'interpretazione geometrica), dove la componente deterministica si riduce al singolo parametro μ.

È possibile estendere la (1) al caso di componenti deterministiche più complesse, costruendo un'intera classe di modelli noti come "modelli lineari".

È importante sottolineare che la (1) è solo un'ipotesi teorica sul processo generatore dei dati, che è assunto noto a meno dei due parametri incogniti μ e σ^2. È possibile, però, stimare questi parametri usando rispettivamente la media e la varianza campionaria calcolate attraverso i dati campionari osservati.

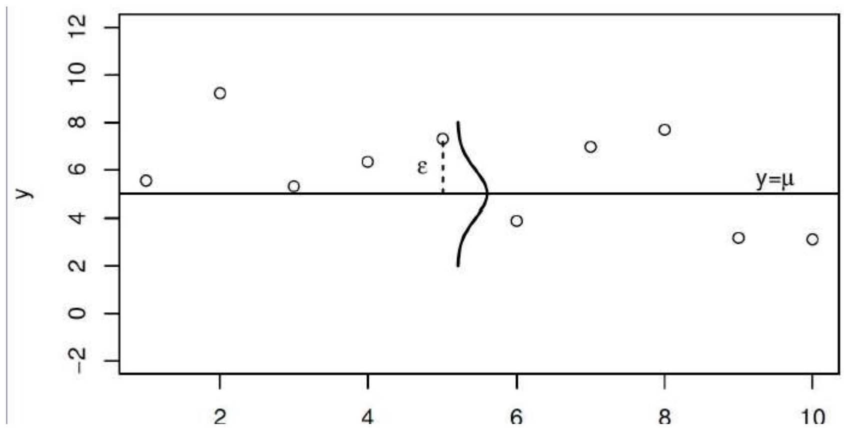

Figura 1: interpretazione geometrica della media aritmetica come modello lineare.

3. *Il modello di regressione univariata*

Supponiamo ora che ogni sia associata ad un valore numerico y_i, che indica una caratteristica dell'unità i-ma (detta 'covariata'). Il modello di regressione univariata postula che la componente deterministica di ogni osservazione sia una funzione lineare della covariata

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (2)$$

Per le proprietà della distribuzione normale, la (2) è equivalente ad assumere che

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

cioè, che le osservazioni siano estratte da normali omoschedastiche (cioè, con la stessa varianza), e con medie che sono una trasformazione lineare della covariata. Quando $\beta_1 = 0$, il modello si riduce al modello (1).

L'interpretazione geometrica della regressione univariata (2) dipende dalla natura della covariata stessa. Se, ad esempio, la covariata è quantitativa, allora β_0 e β_1 sono rispettivamente l'intercetta e il coefficiente angolare di una retta (Figura 2).

Tuttavia, se la covariata è una qualitativa dicotomica, allora l'interpretazione geometrica del modello (2) cambia radicalmente e β_1 indica una differenza tra intercette (Figura 3).

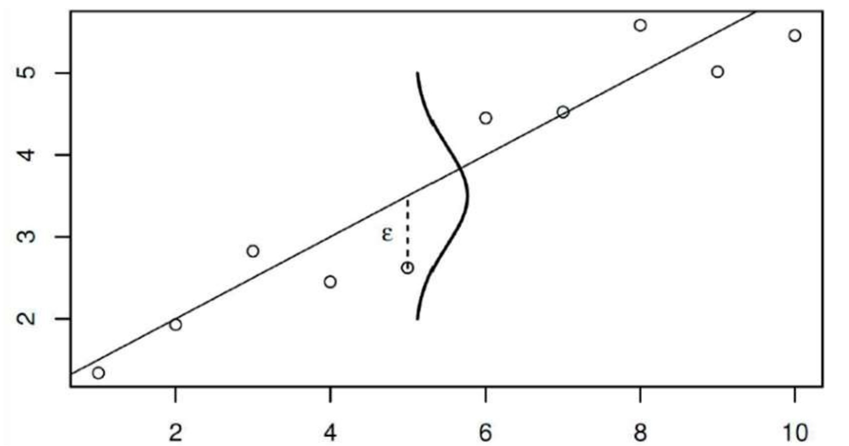

Figura 2: interpretazione geometrica del modello di regressione univariata nel caso di una covariata quantitativa.

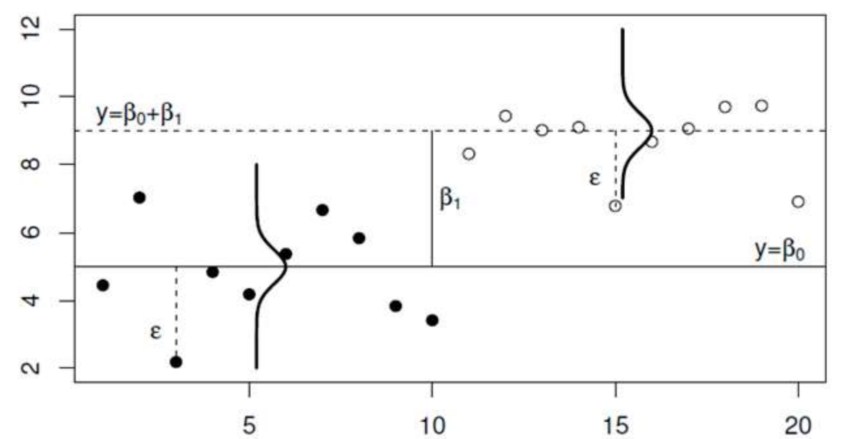

Figura 3: interpretazione geometrica del modello di regressione univariata nel caso di una covariata qualitativa dicotomica.

4. *Il modello lineare*

I modelli lineari estendono il modello di regressione univariata (2) assumendo che la componente deterministica sia una combinazione lineare molteplici covariate:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i,K} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \quad (3)$$

Benché rappresentabili con un'unica equazione, i modelli lineari assumono significati interpretativi assai differenti secondo la natura delle covariate coinvolte. Qui di seguito, elenchiamo i casi più importanti:

1. modelli ANOVA (ANalysis Of VAriance): si tratta di modelli lineari in cui tutte le covariate sono di tipo qualitativo; i coefficienti $\beta_1 \ldots \beta_k$ sono differenze tra intercette;
2. modelli di regressione: sono modelli lineari in cui tutte le covariate sono quantitative; in questo caso i parametri $\beta_1 \ldots \beta_k$ sono i coefficienti angolari di un iperpiano multidimensionale:
3. modelli ANCOVA (ANalysis of COVAriance): sono modelli lineari che contengono alcune covariate qualitative e alcune covariate quantitative.

Un aspetto interessante dei modelli lineari è che può ammettere l'esistenza di covariate ottenute come funzioni di altre covariate. Questa opportunità dà luogo a modelli lineari con interazione e modelli linearizzabili.

I modelli con termini di interazione contengono delle covariate che sono ottenute moltiplicando due covariate. Particolarmente interessanti sono i modelli che contengono l'interazione tra una variabile quantitativa e una qualitativa: in questo caso il coefficiente di regressione associato può essere interpretato come differenza tra due coefficienti angolari.

I modelli linearizzabili sono modelli non lineari che possono essere trasformati in modelli lineari. Un esempio è dato da modelli moltiplicativi del tipo

$$\mu_i = \beta_0 \times x_{i1}^{\beta_1} \times x_{i2}^{\beta_2}$$

Mediante una semplice trasformazione logaritmica, tali modelli possono essere trasformati in modelli lineari

$$log\, \mu_i = log\, \beta_0 + \beta_1\, log\, x_{i1} + \beta_2\, log\, x_{i2}$$

5. *Un'applicazione al riscaldamento globale*

Allo scopo di descrivere la flessibilità dei modelli lineari nelle applicazioni di monitoraggio ambientale, è utile considerare un'applicazione pratica.

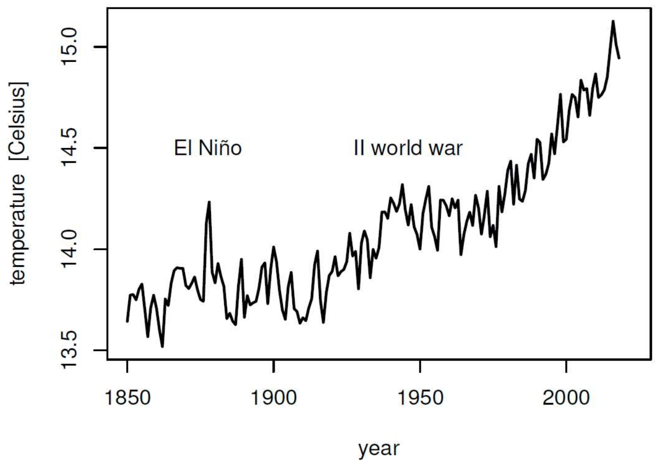

Figura 4: Serie storica dei livelli di temperatura globale. Fonte: Berkeley Earth (http://BerkeleyEarth.org)

La Figura 4 mostra un aumento complessivo della temperatura di circa 1,5 gradi Celsius dal 1850. Il periodo 1850-1940 è stato contrassegnato da un graduale aumento della temperatura, momentaneamente interrotto da un ventennio (1945-1964) in cui le temperature sembrano essersi mantenute su livelli costanti, e seguito da un periodo (dopo il 1965) caratterizzato da un rapido aumento della temperatura. Esistono fluttuazioni annuali minori, ma è difficile determinare quali di queste siano fluttuazioni reali e quali errori di misurazione.

Alcune oscillazioni significative da questi *trend* possono essere facilmente interpretate. Ad esempio, il 1877 e il 1878 hanno temperature notevolmente elevate durante un importante episodio di El Niño in quello che fu soprannominato "l'anno senza inverno". Al contrario, alcuni anni sono caratterizzati da improvvisi cali di temperatura che coincidono con importanti eruzioni vulcaniche.

La Figura 5 illustra, invece, i livelli di anidride carbonica atmosferica, uno dei principali gas serra. I gas serra consentono all'energia solare ad alta frequenza di entrare nella troposfera, bloccando la fuoriuscita di gran parte dell'energia termica a bassa frequenza. C'è un generale consenso nella comunità scientifica sul fatto che l'aumento della temperatura sia dovuto principalmente all'aumento dei gas serra. L'anidride carbonica atmosferica è aumentata di poco più di un terzo dal 1850.

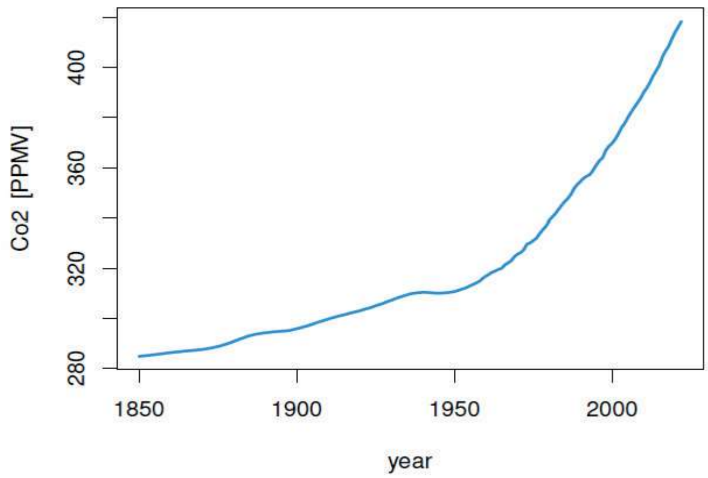

Figura 5: serie storica dei livelli globali di anidride carbonica.

La figura 6 mostra altre due covariate che potrebbero essere coinvolte nella spiegazione del riscaldamento globale: l'irraggiamento solare totale (*total solar iorradiance*, TSI) e il SAOD (*Stratospheric Aerosol Optical Dimming*) che misura il grado di oscuramento solare causato dal particolato alto nella stratosfera.

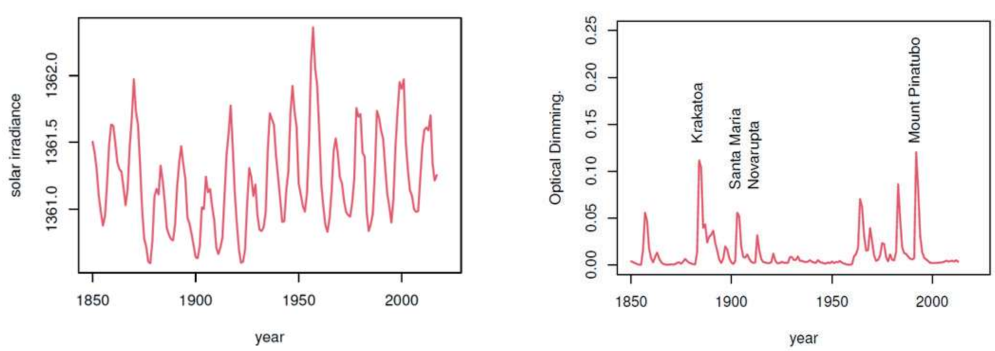

Figura 6: serie storica dell'irraggiamento solare totale (*total solar irradiance*, TSI) e del SAOD (*Stratospheric Aerosol Optical Dimming*), che misura il grado di oscuramento solare causato dal particolato alto nella stratosfera.

Come si nota dalla figura, la covariata TSI ha un ciclo di circa 11 anni che coincide con gli allineamenti planetari e potrebbe giocare un ruolo nella spiegazione delle oscillazioni cicliche delle temperature. Importante è anche la variabile SAOAD. Un maggiore oscuramento porta al raffreddamento riducendo la luce solare che raggiunge la superficie terrestre. Questi dati di oscuramento hanno diversi picchi che coincidono con le principali eruzioni vulcaniche che si sono avute nel periodo oggetto di studio.

Per comprendere l'influenza di queste covariate, possiamo impiegare un modelle statistico lineare del tipo:

$$temperatura = \beta_0 + \beta_1 CO2 + \beta_2 TSI + \beta_3 \mathrm{SAOD} + \beta_4 \text{ El Niño} + \varepsilon \qquad (4)$$

L'equazione (4) include tutte le covariate quantitative illustrate e un'ultima covariata qualitativa (El Niño) che vale 1 in corrispondenza dell'evento estremo e 0 altrimenti. Il modello è dunque un modello di tipo ANOVA senza interazioni.

È possibile stimare i valori dei coefficienti del modello mediante il metodo dei minimi quadrati, ottenendo i risultati della tabella 1.

Tabella 1: stima dei minimi quadrati del modello lineare (4)

	Stima	*p-value*
Intercetta	-50.02	0.17
CO2	0.01	0.00
TSI	0.04	0.09
SAOD	-1.93	0.00
EL NIŇO	0.40	0.00

La seconda colonna della tabella indica quanto l'incremento di una unità di ciascuna variabile contribuisca all'aumento (se il valore è positivo) o alla diminuzione (se il valore è negativo) della temperatura. Ad esempio, ogni parte per milione di anidride carbonica contribuisce a un aumento stimato di 0,01 delle temperature in gradi Celsius. Dal 1850, l'anidride carbonica è aumentata di 128,3 parti per milione, contribuendo con circa 1,283 (= 0,01 x 128,3) gradi all'aumento della temperatura.

La terza colonna della tabella indica invece i *p-value* relativi alle stime ottenute. Un *p-value* superiore a 0.05 indica che la stima ottenuta non è sufficientemente diversa da zero e che quindi la covariata associata non ha un'influenza significativa sulla temperatura. Ad esempio, la variabile TSI non sembra essere significativa. Il risultato è probabilmente dovuto al fatto che le oscillazioni di TSI sono troppo modeste per poter giocare un ruolo importante nell'innalzamento della temperatura.

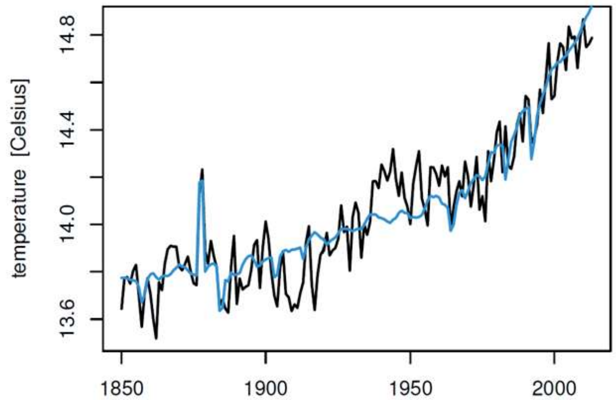

Figura 7: Linea blu: valori predetti dal modello lineare (4) sovrapposti ai livelli osservati di temperatura globale (linea nera).

La figura 7 mostra i valori predetti dal modello (4). Il modello ben sintetizza l'andamento della serie storica della temperatura globale, con l'ausilio di soli 5 parametri. Ma il risultato più importante è che l'aumento di anidride carbonica resta significativo anche dopo aver incluso variabili ambientali importanti che quindi non sono le uniche responsabili del riscaldamento globale. Non abbiamo il potere di controllare l'allineamento planetario o le eruzioni vulcaniche e queste avranno sempre un effetto sulla temperatura. Possiamo però cercare di diminuire le emissioni di anidride carbonica, migliorando la sostenibilità dei nostri processi industriali.

6. *Discussione*

La costruzione di un modello statistico offre valide opportunità di analisi in applicazioni di monitoraggio ambientale, quando le osservazioni sono non sperimentali. In questa sede abbiamo brevemente descritto la classe dei modelli lineari. Benché semplici e facili da interpretare, i modelli lineari offrono la possibilità di valutare l'impatto di variabili quantitative e qualitative sui livelli di una variabile quantitativa. Una semplice applicazione all'importante caso del riscaldamento globale mostra la capacità di sintesi e l'efficacia interpretativa di questa classe di modelli.

Naturalmente i modelli lineari sono solo il primo tassello di una vasta varietà di modelli statistici che via via sono stati introdotti in letteratura e che formano oggi il corpo della Statistica Ambientale.

Suggerimenti bibliografici

MacFarling Meure, C., E1theridge, D., Trudinger, C., Steele, P., Langenfelds, R., van Ommen, T., Smith, A., and Elkins, J. , *Law Dome CO2, CH4 and N2O ice core records extended to 2000 years BP*, in: Geophysical Research Letters, n. 33: L14810, 2021 doi:10.1029/2006GL026152.

Miller, R. L., et al., *CMIP5 historical simulations (1850–2012) with GISS ModelE2*, in «Journal of Advances in Modelling Earth Systems», n. 6, 2014, pp. 441– 478, doi:10.1002/2013MS000266.

Rohde, R. A. and Hausfather, Z. *The Berkeley Earth Land/Ocean Temperature Record*, in «Earth System Scientific Data», n. 12, 2020, pp. 3469–3479, doi:10.5194/essd-12-3469-2020.

Sato, M., Hansen, J. E., McCormick, M. P., and Pollack, J. B. , Stratospheric aerosol optical depths, 1850–1990, in «Journal of Geophysical Research», n. 98(D12), 1993, pp: 22987– 22994, doi:10.1029/93JD02553.

Appendice. Le fonti dei dati

I dati di temperatura sono resi disponibili da Berkeley Earth (http://BerkeleyEarth.org). Berkley Earth è un ente "Indipendente, non governativo e open source" ed è stato originariamente istituito per esaminare la fondatezza di alcune preoccupazioni dei cosiddetti "scettici del clima".

I dati sull'anidride carbonica successivi al 1959 si basano sulle letture operate dall'Osservatorio di Mauna Loa alle Hawaii, e rese disponibili dal Global Monitoring Laboratory (https://www.esrl.noaa.gov/gmd/). I dati storici sull'anidride carbonica si basano invece su numerose letture di carote di ghiaccio profondo polare (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/ice-core).

I recenti dati TSI relativi al periodo 1978-2020 si basano su letture satellitari e sono disponibili sul sito www.ncdc.noaa.gov/cdr/atmospheric/total-solar-irradiance. Altri dati TSI sono stati recuperati da https://data.giss.nasa.gov/modelforce/solar.irradiance.

I dati sull'attenuazione solare dal 1850 al 2012 sono disponibili sul sito: http://data.giss.nasa.gov/modelforce/strataer/ e su https://volcano.si.edu/search_eruption.cfm.