INDICATORI CLIMATICI: I CONTROLLI DI VALIDITÀ E LA RICERCA DEI VALORI ERRATI

F. BAFFO, B. SUATONI, F. DESIATO

Agenzia per la protezione dell'ambiente e per i servizi tecnici (APAT)

Riassunto: La base informativa del Sistema nazionale per la raccolta, elaborazione e diffusione di dati Climatologici SCIA, costituita dai valori statistici decadali, mensili e annuali delle principali variabili misurate dalle reti di osservazione meteoclimatica operative sul territorio italiano, deve essere sottoposta a controlli di validità al fine di assicurare, per quanto possibile, la qualità degli indicatori che vengono elaborati dal sistema. Tali controlli vengono effettuati attraverso un criterio oggettivo di ricerca di valori sospetti o palesemente errati (*outlier*) che, applicato all'insieme degli indicatori derivati dalle osservazioni pluridecennali di temperatura minima e massima giornaliera di 106 stazioni sinottiche, ha portato a rilevare 1334 indicatori *outlier* di cui il 27% è risultato essere derivato da dati di ingresso errati.

Abstract: Ten-days, monthly and annual statistics of the main variables measured by Italian meteorological networks, representing the data core of the national system for the collection, elaboration and diffusion of climatological data SCIA, need to undergo validity controls in order to assure, as much as possible, their quality and reliability. The controls are performed through an objective criterion for searching suspicious or clearly wrong values (outliers) that we applied to temperature indicators derived from multi-decadal minimum and maximum daily temperature data of 106 synoptic stations. The results are that, 27% of the 1334 found outliers, may be attributed to mistakes in input data series and have been corrected or eliminated from the database.

Introduzione

L'Agenzia per la protezione dell'ambiente e per i servizi tecnici (APAT) ha realizzato di recente, in collaborazione con l'Ufficio Generale per la Meteorologia dell'Aeronautica Militare (UGM), l'Ufficio Centrale di Ecologia Agraria (UCEA) del Consiglio per la Ricerca e Sperimentazione per l'Agricoltura e il Servizio Idrometeorologico Regionale dell'ARPA Emilia Romagna (SIM), un sistema informatizzato per la raccolta, elaborazione e diffusione di dati meteoclimatici, denominato SCIA. Lo scopo principale del sistema è di elaborare e mettere a disposizione i valori statistici (denominati convenzionalmente "indicatori") con aggregazione temporale decadale, mensile e annuale, derivati dalle serie temporali delle variabili meteorologiche misurate da diverse reti di osservazione, e aggiornare periodicamente le informazioni con una procedura standardizzata. Attraverso gli indicatori vengono elaborati anche i valori climatologici normali su un periodo standard di 30 anni (il più recente, come indicato dall'Organizzazione Meteorologica Mondiale, è il trentennio 1961-1990) e i valori di anomalia, cioè delle differenze tra i valori di un determinato anno e i valori normali.

Le informazioni prodotte da SCIA sono accessibili attraverso un sito web dedicato, all'interno del portale (SINAnet) del sistema nazionale conoscitivo e di informazione ambientale, all'indirizzo www.scia.sinanet.apat.it. Attraverso il sito web di SCIA è possibile visualizzare sotto forma di tabelle, grafici e mappe, e scaricare su file di testo, i principali indicatori, i valori normali e le anomalie elaborati e memorizzati dal sistema.

Le principali variabili meteoclimatiche trattate da SCIA sono: temperatura, temperatura potenziale, temperatura equivalente potenziale, precipitazioni, umidità relativa, vento, bilancio idrico, indici bioclimatologici, eliofania, evapotraspirazione, gradi giorno, nebbia e visibilità, nuvolosità, pressione atmosferica, radiazione globale.

Per ciascuna variabile viene calcolato l'insieme degli indicatori rappresentativi del fenomeno climatico ad essa associato e della sua distribuzione statistica. Per esempio, per le precipitazioni vengono calcolati il valore cumulato, il valore massimo, la data di occorrenza del massimo, la distribuzione dei valori di precipitazione cumulata su 1, 6, 12 e 24 ore, il numero di giorni con neve, il numero di giorni con temporale e il numero di eventi temporaleschi.

Le serie di indicatori statistici decadali, mensili ed annuali sono calcolate a partire da dati meteorologici elementari, diversi per contenuto e per formato a seconda della fonte di provenienza. Il contenuto varia in funzione della tipologia della stazione di misura (automatica o manuale), della strumentazione e della modalità d'acquisizione e archiviazione dei dati. In particolare, possono essere diversi la frequenza delle osservazioni e il tempo di media del dato rilevato. Ad esempio, per la rete UGM vengono elaborate le osservazioni sinottiche alla superficie (messaggi triorari SYNOP e riepilogativi giornalieri SYREP), mentre per l'UCEA vengono elaborati i dati rilevati dagli osservatori (tre al giorno), dalle stazioni termopluviometriche (un dato giornaliero) e dalle stazioni automatiche delle Rete Agrometeorologica Nazionale (dati orari). I criteri generali adottati per il calcolo degli indicatori sono definiti dall'Organizzazione Meteorologica Mondiale (WMO, 1990).

Un aspetto di grande rilievo ai fini dell'utilizzo e della interpretazione degli indicatori climatici prodotti da SCIA, è costituito dalla qualità degli indicatori stessi, che dipende ovviamente, oltre che dalla solidità della procedura di calcolo, dalla qualità dei

dati meteorologici di ingresso. Nel presente lavoro sono descritti l'insieme dei controlli di validità cui sono sottoposti gli indicatori climatici, i criteri e i metodi di ricerca dei valori errati, e i risultati della loro applicazione agli indicatori di temperatura finora elaborati e presenti sul database di SCIA.

I controlli di validità

Ciascun indicatore elaborato attraverso SCIA è contrassegnato da un *flag* di validità. In generale, il criterio per l'attribuzione del *flag* di indicatore valido consiste nella disponibilità di almeno il 75% di dati elementari validi che concorrono al calcolo dell'indicatore stesso. Questo criterio, che tiene conto della presenza di dati mancanti distribuiti nelle serie in modo generalmente irregolare, rappresenta un compromesso tra l'esigenza di non scartare un numero elevato di dati utili e la necessità di ottenere indicatori sufficientemente rappresentativi nell'intervallo di tempo considerato. Per i valori di precipitazione cumulata che, per la natura stessa dell'indicatore, dovrebbero essere considerati validi soltanto se sono disponibili tutti i dati di origine, la soglia di dati disponibili e validi oltre la quale è assegnato il *flag* di indicatore valido è il 90%. Accanto al valore di ciascun indicatore viene comunque calcolato e conservato il numero di dati che lo ha generato.

I controlli di validità dei dati di ingresso, cioè delle serie temporali delle osservazioni meteoclimatiche disponibili per ciascuna rete ricadono, in generale, sotto la titolarità e la responsabilità delle fonti. Pertanto, si richiede che i dati siano validi, oppure che siano contrassegnati da un *flag* di validità (uno per dato) presente sui file di ingresso. La procedura di calcolo di SCIA prende in considerazione, sia per la determinazione del valore dell'indicatore, sia per il conteggio dei dati utili alla sua determinazione, solo i dati con *flag* di dato valido. Per i dati sinottici della rete UGM vengono inoltre utilizzate e applicate sequenzialmente due classi di controlli: un controllo climatologico debole ed un controllo di consistenza interna. Il controllo climatologico debole si basa sul requisito di non superamento di soglie minime e massime di accettazione del dato, abbastanza blande. Le soglie sono state definite a priori utilizzando dei criteri di ragionevolezza e non derivano pertanto da un'analisi statistica preventiva della base dati disponibile. Il controllo di consistenza interna prevede il controllo reciproco di più variabili in uno stesso istante temporale. Ad esempio, un controllo di tale tipo, applicato alla temperatura dell'aria, la mette a confronto con la temperatura di rugiada: non deve mai verificarsi che la temperatura dell'aria sia inferiore alla temperatura di rugiada.

Uno o più dati errati all'origine e che (nel caso della rete UGM) abbiano comunque superato il controllo climatologico debole e il controllo di consistenza interna, generano inevitabilmente valori errati degli indicatori decadali, mensili e annuali. L'entità e l'evidenza dell'errore dipendono dall'entità dell'errore del dato originale e dal tipo di indicatore. In generale, un dato errato risulta poco visibile dall'analisi dei valori medi annuali, mentre può essere ben evidenziato dall'analisi delle serie di valori estremi. È in ogni modo molto importante definire e sperimentare una metodologia efficace di ricerca e d'individuazione degli indicatori non validi e, successivamente, provvedere alla loro correzione o eliminazione dalla base informativa di SCIA. In questo modo, tra l'altro, si assicura una migliore qualità dei valori climatologici normali, delle anomalie, e delle mappe relative alle principali variabili climatiche, che si ottengono dalla interpolazione spaziale degli indicatori validi.

La Ricerca dei valori errati

Teoria

In generale, i metodi finalizzati alla ricerca dei valori errati all'interno di una serie climatologica di dati (Pavan et al., 2003; Ridley et al., 2004) sono basati sul concetto di *outlier*, cioè di dati che non rientrano in un determinato *range* di valori, delimitato da particolari soglie climatologiche. Per rendere la ricerca e l'individuazione di *outlier* il più possibile automatica e oggettiva sono state sviluppate alcune metodologie (Gandin, 1963) (StatSoft, Sito Internet) basate su diversi " test di consistenza" così classificati (Plummer et al., 2003):
test di consistenza interna, basati sul confronto di due variabili in uno stesso istante;
test di consistenza temporale, che analizzano la persistenza nel tempo dei diversi elementi climatologici;
test di consistenza spaziale, che considerano la naturale variazione nello spazio delle variabili climatiche.
In particolare, se si considera il mese come base di aggregazione dei dati, i test di consistenza temporale partono dall'assunzione che il singolo valore mensile di una serie dovrebbe essere congruente con i valori dello stesso mese di tutti gli altri anni della serie stessa (Eisched et al., 1995), il che si traduce nell'individuazione dei limiti oltre i quali il dato non è considerato valido. I valori limite sono calcolati in funzione del *range* interquartile (IR) della distribuzione dei dati di ogni stazione e ogni mese dell'anno. Specificando meglio, un valore è considerato *outlier* quando:

$$X_i - q50 > fIR$$

dove X_i è il valore medio mensile dell'anno i, $q50$ è la mediana, IR è il range interquartile definito come differenza tra il 75° e il 25° percentile e f è un fattore di moltiplicazione.

Applicazione

Poiché l'applicazione dei metodi di ricerca degli *outlier* a ciascuna serie temporale degli indicatori di tutte le variabili e di tutte le stazioni che sono memorizzati su SCIA è un'operazione al di là delle possibilità pratiche e degli scopi del sistema, l'approccio utilizzato è stato quello di determinare i valori limite che definiscono gli *outlier*, per *cluster* di indicatori composti da elementi-campione con caratteristiche climatologiche simili. Una volta determinati i valori limite per ogni *cluster*, gli stessi dovrebbero essere utilizzati per tutta la base di indicatori del sistema, presente e futura. Il raggruppamento in cluster è stato effettuato con i seguenti criteri:
area geografica: il territorio nazionale è stato suddiviso in 3 macro-regioni: nord, centro, sud e isole;
stagione: sono state considerate tre classi:
inverno (dicembre – gennaio – febbraio – marzo)
estate (giugno – luglio – agosto – settembre)
primavera-autunno (aprile – maggio – ottobre – novembre)

quota sul livello del mare (suddivisione in 3 o 4 classi);
vicinanza al mare (suddivisione in 2 classi).

Il campione è costituito da tutti gli indicatori mensili di temperatura delle stazioni della rete sinottica, già presenti su SCIA. Allo scopo di garantirne la significatività statistica, l'analisi è stata circoscritta alle stazioni che presentavano in partenza, per ciascun mese dell'anno, almeno 20 indicatori validi, cioè 20 anni con indicatore valido. In questo modo, il campione è risultato così composto: 38 stazioni per il nord, 27 per il centro e 41 per il sud e le isole.

Per ciascuna macro-regione e per ogni stagione sono stati così calcolati:
la mediana;
il 25° e il 75° percentile;
il range interquartile IR.

Per quanto riguarda la costante *f*, dalla letteratura sono stati rilevati valori compresi tra 2.75 (Eischeid et al., 1995) e 3 (Velleman e Hoaglin, 1981). Il valore di *f* può essere determinato empiricamente individuando su un grafico No(*f*), dove No è il numero di *outlier*, il valore di *f* in corrispondenza del quale la pendenza della curva *p(f)* è prossima a 0. Un grafico di questo tipo, per i valori medi della temperatura massima giornaliera, è presentato in fig. 1. Nel presente lavoro è stato scelto infine un valore leggermente più conservativo, $f = 2.5$, che restringe il *range* di indicatori validi ed estende i range di ricerca degli indicatori errati. Pur comportando un maggior onere per la verifica di un maggior numero di *outlier,* questa scelta ha il vantaggio di garantire una ricerca dei valori errati più ampia ed accurata.

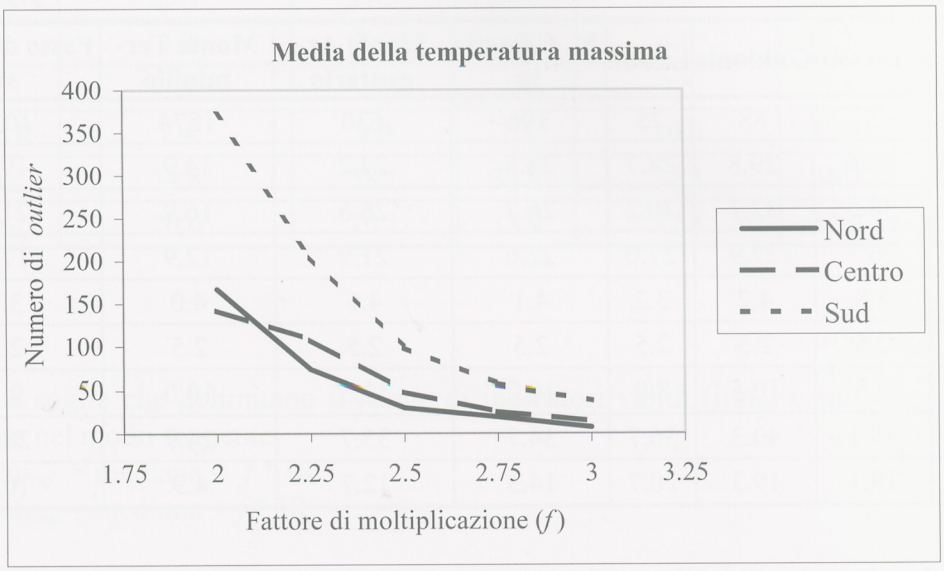

FIG. 1: Media della temperatura massima giornaliera, stagione estiva. Numero di *outlier* rilevati in funzione del fattore di moltiplicazione del *range* interquartile f.

Risultati e Discussione

A titolo di esempio, nelle tabelle 1a, 1b, 1c e 1d sono mostrati i risultati della ricerca di *outlier* del valore medio della temperatura massima giornaliera, per le stazioni dell'Italia centrale, stagione estiva.

TAB. 1a

	Arezzo	Campobasso	Civitavecchia	Falconara	Firenze/ Peretola	Frontone	Frosinone
quota	248	793	3	12	40	570	180
mediana	27.9	24.7	25.8	26.7	29.1	24.7	28.7
75° perc	29.9	26.6	27.0	28.0	30.9	26.9	30.8
25°perc	25.9	22.4	24.4	25.2	27.0	22.7	26.8
IR	4.0	4.2	2.6	2.8	3.9	4.2	4.0
f	2.5	2.5	2.5	2.5	2.5	2.5	2.5
***f** * IR**	10.0	10.5	6.5	7.0	9.7	10.5	10.0
lim_{sup}	37.9	35.2	32.3	33.7	38.8	35.2	38.7
lim_{inf}	17.9	14.2	19.3	19.7	19.4	14.2	18.7

TAB. 1b

	Grosseto	Guidonia	Latina	M. Calamita	Monte Argentario	Monte Terminillo	Passo della Cisa
quota	5	88	25	396	630	1874	1039
mediana	28.6	29.8	28.7	24.5	24.2	14.9	19.2
75° perc	30.5	32.1	30.2	26.7	26.5	16.9	21.1
25°perc	26.7	27.9	27.0	22.6	21.9	12.9	17.4
IR	3.8	4.2	3.2	4.1	4.6	4.0	3.7
f	2.5	2.5	2.5	2.5	2.5	2.5	2.5
***f** * IR**	9.5	10.5	8.0	10.2	11.5	10.0	9.1
lim_{sup}	38.1	40.3	36.7	34.7	35.7	24.9	28.3
lim_{inf}	19.1	19.3	20.7	14.3	12.7	4.9	10.1

TAB. 1c

	Perugia	Pescara	Pisa/ S. Giusto	Ponza	Pratica di mare	Radicofani	Roma/ Ciampino
quota	208	10	2	184	6	816	129
mediana	27.7	27.3	27.7	25.9	26.8	23.1	28.9
75° perc	29.9	28.8	29.1	27.4	28.3	25.4	30.6
25°perc	25.8	25.8	25.9	24.3	25.5	21.0	27.0
IR	4.1	3.0	3.2	3.1	2.8	4.4	3.6
f	2.5	2.5	2.5	2.5	2.5	2.5	2.5
***f** * IR**	10.3	7.6	7.9	7.6	7.0	10.9	9.0
lim_{sup}	38.0	34.9	35.6	33.5	33.8	34.0	37.9
lim_{inf}	17.5	19.7	19.7	18.3	19.8	12.1	19.9

TAB. 1d

	Roma/Fiumicino	Roma/Urbe	Termoli	Vigna di Valle	Viterbo	Volterra
quota	2	18	16	262	300	555
mediana	27.4	29.1	26.1	27.3	27.5	23.7
75° perc	28.8	31.1	27.5	29.3	29.5	25.7
25°perc	25.9	27.5	24.6	25.3	25.3	21.7
IR	3.0	3.6	2.9	4.0	4.2	4.0
f	2.5	2.5	2.5	2.5	2.5	2.5
***f** * IR**	7.4	9.0	7.3	10.0	10.5	10.0
lim_{sup}	34.8	38.1	33.4	37.3	38.0	33.7
lim_{inf}	20.0	20.1	18.9	17.3	17.0	13.7

Le soglie che delimitano il *range* di indicatori validi (lim_{inf} e lim_{sup}) sono state calcolate nel modo seguente:

lim_{inf} = mediana – f * IR

lim_{sup} = mediana + f * IR

Successivamente, sono stati rappresentati graficamente i punti di coordinate (lim_{inf}, lim_{sup}) rappresentativi di ciascuna stazione, ottenendo una nube di punti che mostra con maggiore o minore evidenza, caso per caso, i *cluster* di stazioni con caratteristiche climatiche simili (v. per esempio fig. 2).

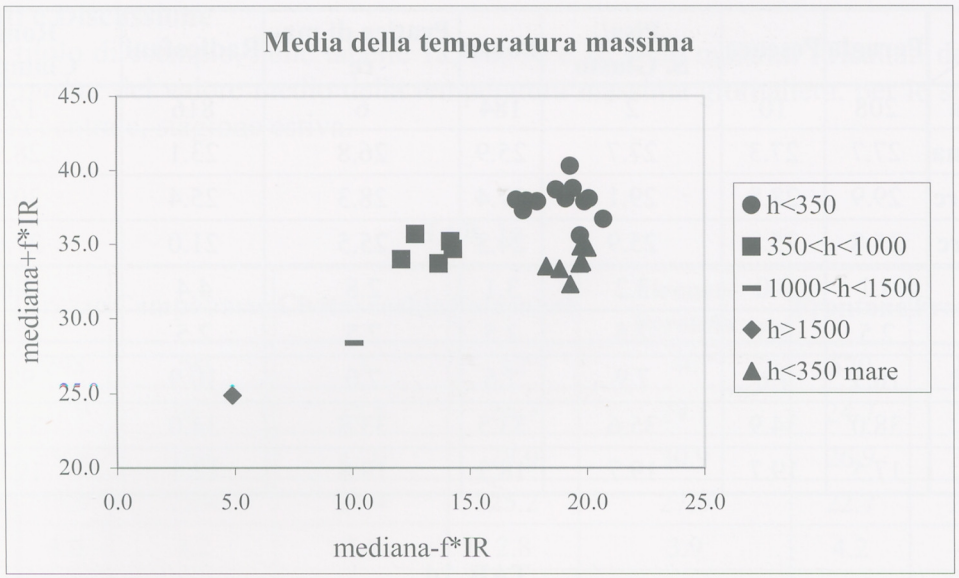

FIG. 2: Limiti inferiore (asse x) e superiore (asse y) del *range* di validità della media della temperatura massima giornaliera, per diverse classi di altitudine h. Stazioni dell'Italia centrale, stagione estiva.

Infine, le soglie di ciascun *cluster* sono state calcolate come media aritmetica dei limiti di ciascuna stazione appartenente a quel *cluster*.

Una volta determinati i valori soglia per ogni cluster, è stata effettuata la ricerca degli *outlier* su tutti gli indicatori mensili delle seguenti variabili:

Temperatura massima: media e valore massimo;

Temperatura minima: media e valore minimo;

Temperatura media, calcolata come media delle temperature massima e minima.

Per tutti gli *outlier* trovati, è stata effettuata una verifica puntuale di quali e quanti di essi fossero effettivamente indicatori errati, cioè fossero determinati da uno o più dati di origine errati, e quali e quanti fossero, al contrario, indicatori validi seppure con valori al di fuori del *range* calcolato per il *cluster* corrispondente. Tale verifica si basa sull'analisi delle serie di dati originali da cui è derivato ciascun *outlier*, operando un controllo sulla continuità temporale della serie e/o sulla correlazione spaziale con dati di stazioni limitrofe.

I risultati sono sintetizzati in tabella 3, dove viene riportato, per ogni indicatore, macro-regione e stagione, il numero di indicatori errati rispetto al numero di *outlier*.

TAB. 3a

Tmax assoluta	Inverno	Estate	Primavera/Autunno
Nord	**23 / 27**	**2 / 5**	**5 / 7**
Centro	**6 / 59**	**1 / 3**	**1 / 1**
Sud	**12 / 104**	**5 / 35**	**2 / 6**

TAB. 3b

Tmin assoluta	Inverno	Estate	Primavera/Autunno
Nord	**6 / 53**	**51 / 80**	**15 / 17**
Centro	**11 / 96**	**28 / 50**	**7 / 10**
Sud	**28 / 93**	**54 / 65**	**30 / 42**

TAB. 3c

Tmax media	Inverno	Estate	Primavera/Autunno
Nord	**13 / 23**	**0 / 7**	**0 / 0**
Centro	**0 / 47**	**0 / 0**	**1 / 1**
Sud	**2 / 91**	**4 / 7**	**0 / 0**

TAB. 3d

Tmin media	Inverno	Estate	Primavera/Autunno
Nord	**1 / 26**	**0 / 22**	**1 / 1**
Centro	**0 / 53**	**0 / 13**	**0 / 3**
Sud	**9 / 143**	**10 / 15**	**7 / 11**

TAB. 3e

Tmedia	Inverno	Estate	Primavera/Autunno
Nord	**1 / 9**	**0 / 7**	**0 / 0**
Centro	**0 / 28**	**0 / 0**	**0 / 0**
Sud	**7 / 43**	**10 / 11**	**7 / 20**

Complessivamente, sono stati individuati 1334 *outlier*, di cui il 27 % è risultato corrispondere a valori effettivamente errati degli indicatori, cioè a indicatori derivati da dati errati all'origine. Come era prevedibile, il numero di *outlier* è decisamente più elevato per gli indicatori relativi ai valori estremi, rispetto ai valori medi. Il numero di *outlier* e la percentuale di indicatori errati variano sensibilmente da un *cluster* all'altro; ciò può essere dovuto sia a differenze naturali tra le distribuzioni statistiche degli indicatori nei diversi *cluster*, sia al fatto che l'algoritmo di calcolo dei limiti inferiore e superiore non è eviden-

temente tarato e ottimizzato per ciascun *cluster*. D'altra parte, va ricordato che lo scopo di questa analisi è la definizione di una procedura semiautomatica che consenta di individuare e di eliminare o correggere il maggior numero di indicatori errati, limitando però per quanto possibile l'onere di riesaminare dettagliatamente le serie di dati di origine con valori "sospetti". In altre parole, l'*optimum* è costituito dall'equilibrio tra una ricerca di valori errati troppo vasta (che corrisponde a un *range* ristretto di valori ritenuti validi, e a un numero elevato di *outlier*), e una troppo ristretta (che corrisponde a un *range* allargato di valori ritenuti validi, e a un numero ridotto di *outlier*).

Conclusioni

Lo scopo di questo studio è di definire un procedimento oggettivo per la ricerca e l'individuazione di indicatori climatici errati, perché derivati da uno o più dati errati sulle serie temporali di ingresso. Una volta definito l'algoritmo per la determinazione delle soglie di validità degli indicatori di temperatura, esso è stato applicato alle stazioni sinottiche con almeno 20 anni di dati, e i risultati sono stati raggruppati in *cluster* di stazioni con caratteristiche climatiche simili. Applicando i valori soglia rappresentativi di ciascun *cluster,* sono stati individuati complessivamente 1334 *outlier*, e una successiva verifica puntuale degli errori presenti sulle serie di dati meteo di ingresso ha consentito di individuare 360 indicatori errati, di cui 287 relativi alle temperature estreme. L'applicazione sistematica di un procedimento di ricerca degli *outlier* e di determinazione degli indicatori errati, del tipo di quella descritta, dovrebbe consentire di migliorare significativamente la qualità delle informazioni messe a disposizione dal sistema SCIA, e in particolare di tutte le elaborazioni di sintesi, come le tabelle e le mappe dei valori climatologici normali e delle anomalie, prodotte con programmi che attingono dinamicamente alla base di indicatori contrassegnati dal corretto *flag* di validità.

Bibliografia

EISCHED J.K., BRUCE BAKER C., KARL T.R., DIAZ H.F., 1995: *The quality control of long-term climatological data using objective data analysis*, "Journal of Applied Meteorology", 34, pp. 2787-2795.

GANDIN, L.S., 1963: *Objective analysis of meteorological fields*, "Gidrometeorologicheskoe Izdatel'stvo", Leningrad.

PAVAN, V., TOMOZEIU, R., SELLINI, A., MARCHESI, S., MARSIGLI, C., 2003: *Controllo di qualità dei dati giornalieri di temperatura minima e massima e di precipitazione*, "Quaderno Tecnico ARPA-SIM", 15.

PLUMMER, N., ALLSOPP, T., LOPEZ, J. A., 2003: *Guidelines on Climate Observation Networks and Systems*, "WMO/TD", 1185.

RIDLEY B., BOLAND J., LUTHER M., 2004: *Quality control of Climate Data Sets. Solar 2004: Life, the Universe and Renewables*.

http://www.statsoft.com/support/faq5/stbasic/outliers.html

VELLEMAN P.F., HOAGLIN D.C., 1981: *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press.

WMO (WORLD METEOROLOGICAL ORGANIZATION), *1990*, "Guide to climatological practices", seconda edizione, Ginevra, Svizzera *(alcuni capitoli di una edizione successiva non ancora pubblicata sono reperibili al sito web del WMO).*