A quantile matching adjustment algorithm for Gaussian data series

Xiaolan L. Wang

Climate Research Division, Science and Technology Branch

Environment Canada, Toronto, Canada

December 3, 2009

The Quantile Matching (QM) algorithm described below is similar to that proposed in Wang et al. (2009). The objective of the QM adjustments is to adjust the base series so that the empirical distributions of all segments of the detrended base series match each other. The adjustment value depends on the empirical frequency of the datum to be adjusted (i.e. it varies from one datum to another in the same segment, depending on their corresponding empirical frequencies).

Let $\{Y_i,\ i = 1, \ldots, N\}$ denote a data series observed at times $t_1 < \cdots < t_i < \cdots < t_N$, which consists of the annual cycle C_m, a common linear trend component βt_i (β could be zero), and identically distributed Gaussian random variations. Assuming that a list of changepoints have been identified for series $\{Y_i\}$, e.g. by using the PMFred or PMTred algorithm of Wang (2008). Importantly, in the PMFred or PMTred algorithm, the annual cycle, lag-1 autocorrelation, and linear trend of the base series were estimated in tandem while accounting for all identified shifts (Wang 2008). One can subtract the estimated linear trend component from series $\{Y_i\}$, obtaining the detrended series $\{X_i = Y_i - \hat{\beta} t_i\}$. In order to preserve the linear trend component in the base series, this detrended series $\{X_i\}$ is then used to estimate the empirical cumulative distribution function (ECDF) for each segment of the base series and the adjustments needed to make the base series homogeneous. These are detailed next.

Let M_q denote the number of points at which the ECDF will be estimated. The

detrended data in each segment is sorted in ascending order and then divided into M_q ascending categories "equally" (to the extent possible). Let $Z_s(l)$ denote the mean of the l-th category of the s-th segment of series $\{X_i\}$, and $F(l)$ the upper bound of the empirical cumulative frequency (ECF) of the data in the l-th category. In other words, all the l-th category data fall within the percentile range $(F(l-1), F(l)]$. The following differences in the category mean are derived:

$$D_s(l) = \overline{Z}_S(l) - \overline{Z}_s(l) \quad (l = 1, 2, ..., M_q)$$

where S denote the segment to which the other segments are to be adjusted (i.e. the base segment). For Gaussian data, let $F(0) = 0$ and $D_s(0) = D_s(1)$; also let $F(M_q + 1) = 1 + 1/M_q$ and $D_s(M_q + 1) = D_s(M_q)$ (these boundary conditions keep the adjustments for the lowest and highest $50/M_q$ percent of data bounded, not to let them depart too much from the mean adjustment for the corresponding category). Thus, for each segment s, there are $(M_q + 2)$ data points, $(F(l) - 0.5/M_q, D_s(l))$ for $l = 0, 1, ..., M_q + 1$ [here, the subtraction of $0.5/M_q$ from the $F(l)$ is to put the $D_s(l)$ at the center ECF of the l-th category, whose ECFs fall within the interval $(F(l-1), F(l)]$]. As shown in Fig. 1, a natural cubic spline is then fitted to these (M_q+2) data points for each segment s [except segment S, for which $D_S(l) \equiv 0$], which will be used to derive the adjustments needed to homogenize the series, as described next.

Let $\mathcal{F}_s(i)$ denote the empirical cumulative frequency of the ith datum in segment s of series $\{X_i\}$. From the fitted spline, we can look up the difference that corresponds to the frequency $\mathcal{F}_s(i)$ (i.e. use $\mathcal{F}_s(i)$ as the X-axis value in Figure 1 to look up the corresponding Y-axis value). The difference $\mathcal{D}_s(i)$ is the amount that will be added to the ith datum in segment s of series $\{X_i\}$, to adjust it to segment S. This spline interpolation is carried out for each value in each segment except segment S. The resulting differences $\mathcal{D}_s(i)$ for $i = 1, 2, ..., N$, are referred to as the QM adjustments, which vary with the frequency and hence can account for a seasonality of discontinuity if it exists (e.g., it is possible that

winter and summer temperatures are affected differently and are thus adjusted differently, because they belong to the lower and upper quartiles, respectively; see Fig. 1). The QM adjusted base series is obtained by adding the linear trend component $\hat{\beta}t_i$ back to the QM adjusted version of series $\{X_i\}$. Therefore, the trend component estimated for the base series is preserved in the QM adjusted series, which is very important.

The number M_q here shall be determined so that the shortest segment in the series has enough data in each of the M_q categories. Let N_{min} denote the length of the shortest segment in the series. In our software package, the actual choices of M_q value include any integer between 1 and 20 inclusive; users can either set $M_q = 0$ to let the codes determine the appropriate M_q value, or chose a M_q value from any integer between 1 to 20 (any larger number will be replaced by 20 automatically). In order to ensure that even the shortest segment has enough data in each category, a M_q value chosen by a user will be replaced by the integer $N_{min}/5$ (or $N_{min}/20$ for daily data) if the chosen M_q is larger [this ensures that there are at least 5 data (or 20 daily data) in each category for estimating the categorical mean]. If the procedure results in $M_q = 0$ (meaning that there is not enough data in the shortest segment for estimating QM adjustments), our codes re-set $M_q = 1$ to use one single adjustment value for all data in the same segment (i.e., the usual mean adjustment). We do not recommend set M_q larger than 20, because the larger the M_q, the fewer data available for estimating the ECDF and the mean between-segment differences for each category, and hence the larger the sampling variability and uncertainty of the estimates of adjustments. Also note that any $M_q \geq 2$, the resulting adjustment varies from one datum to another (e.g., when $M_q = 2$, not only two different adjustment values are applied to the values in a segment, but each and every datum in a segment has its own adjustment that corresponds to its empirical cumulative frequency), because of the spline fit and interpolation described above.

Note that all quantile matching algorithms (e.g., Della-Marta and Wanner 2006, Trewin

and Trevitt 1996, and the one described above) try to line up the adjustments by empirical frequencies, implicitly assuming that the frequency series are homogeneous. Thus, quantile matching algorithms would work well when the frequency series is homogeneous (e.g. for continuous variables such as temperature data); but they will not work, could even be problematic when a discontinuity is present in the frequency series of a discontinuous variable such as daily precipitation amounts (or wind speeds).

Note that, as a result of applying the QM adjustments with $M_q \geq 2$, the whole distribution of the data, not only the mean, could be adjusted. When the PMFred or PMTred algorithm is used without data transformation, the detection of shifts is done on the mean only, which indicates that any shift that occurs in the variance or higher order statistic without a significant shift in the mean may go undetected in this case.

References

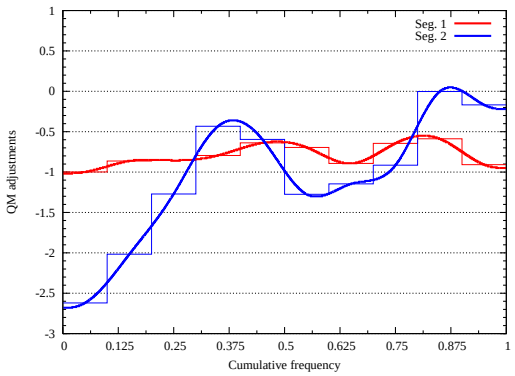
Della-Marta, P. M. and H. Wanner, 2006: A Method of Homogenizing the Extremes and Mean of Daily Temperature Measurements. *J. Climate*, **19**, 4179-4197.

Trewin, B. C. and A. C. F. Trevitt, 1996: The Development of Composite Temperature Records. *Int. J. Climatol.*, **16**, 1227-1242.

Wang, X.L., H. Chen, Y. Wu and Q. Pu, 2009: New techniques for detection and adjustment of shifts in daily precipitation data series. *J. Appl. Meteor. Climatol.* (submitted)

Wang, X.L., 2008: Accounting for Autocorrelation in Detecting Mean Shifts in Climate Data Series Using the Penalized Maximal t or F test. *J. App. Meteor. Climatol.* **47**, 2423-2444. DOI:10.1175/2008JAMC1741.1.

a. distribution of QM adjustments with $M_q = 10$



b. distribution of QM adjustments with $M_q = 20$

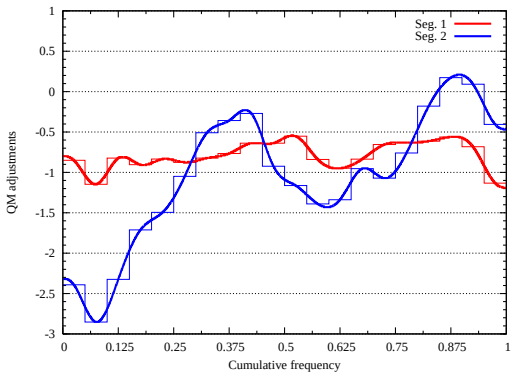


Figure 1. The $F(l) \sim D_s(l)$ relationship for each segment that needs to be adjusted (step lines) and the corresponding fitted natural cubic splines (i.e., distribution of the QM adjustments over empirical cumulative frequency) for two different M_q values.