International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

Quantile regression ensemble for summer temperatures time series and its impact on built environment studies

Manuel Herrera [a], **Matthew Eames** [b], **Alfonso Ramallo-Gonzalez**[a], **Chunde Liu**[a] **and David Coley**[a]

[a] *EDEn - ACE Dept., University of Bath, United Kingdom*
(amhf20@bath.ac.uk, aprg20@bath.ac.uk, c.liu2@bath.ac.uk, d.a.coley@bath.ac.uk)

[b] *CEE - School of Physics, University of Exeter, United Kingdom*
(m.e.eames@exeter.ac.uk)

Abstract: The occurrence of heat waves, heavy rainfall, or drought periods have an increasing trend nowadays as consequence of Earths global warming. This seriously affects natural habitats and directly impacts on the human environment. Architects and engineers use different approaches to model reference and future weather conditions to achieve building designs that resilient and energy efficient. However, modelling extreme weather events within those future conditions is a relative few explored field. This paper introduces the analysis of regression models for summer temperature time series that facilitate the representation of those extreme events; this is based on upper and lower quantiles, instead of the ordinary regression which is conditioned by the mean. The advantage of this proposal is to focus on finding patterns for both higher temperatures during the day and warmer temperatures during the night. These two temperatures are key in the study of overheating and therefore thermal related morbidity as the night time temperature will show the possibility of night purging. Advances on quantile regression models coming from both resampling and ensemble approaches are further investigated and compared with the state of the art. The results are sound in terms of statistical robustness and knowledge of inputs that specifically affect to temperature extremes. The new method has been tested with real data of temperatures in London for a period of 50 years. The quantile regression technique provides more meaningful information for use in building simulation than standard methods.

Keywords: Quantile regression, models ensemble, weather, heat waves, buildings

1 INTRODUCTION

Weather forecasting systems have always had a great importance in society and its advances in recent decades have resulted in numerous benefits for social and economic activities. As weather forecasts warn of extreme conditions in a short time ahead, analysing weather patterns for arranging climate projections is a priority to define prevention policies on adaptation to changes in medium or extreme environmental conditions; ranging from floods and droughts to heat waves. This analysis will become more important as the frequency and intensity of severe weather events increase as a result of the Earth's global warming [IPCC, 2015]. The building sector is not an exception in terms of how it is becoming affected by global warming and extreme events. On the contrary, it has been seen that building are magnifiers of weather events. It is the reason why is common place now to use building simulators with representations of the weather in weather files to evaluate the thermal designs of buildings.

This paper aims to investigate the use of quantile regression analysis with weather patterns of warmer temperatures during the summer time to improve the characterisation of extreme events in weather files. Overheating commonly causes comfort issues associated with reduced productivity at work environment and illness in babies and elderly population [de Wilde and Coley, 2012]. In extreme cases, the consequences of overheating could be a peak of morbidity as in the European heatwave of 2003, where almost 70,000 people died. During a heatwave, temperatures remain high also during the night, preventing a cooling cycle in the building via night purging [Coley and Kershaw, 2010], this has major consequences for the occupants of buildings, as a sustained hot internal environment leads to the inability of occupants to maintain their body temperature. In addition to the already mentioned issues in the built environment [Coley et al., 2012]; the disciplines of agriculture [Wilby et al., 2010], water resources [Christierson et al., 2012; Brentan et al., 2016] and energy systems [Ebinger, 2011] are also affected by weather conditions under climate change. Currently, methods for Building Engineering predicting extreme weather events consists on carrying out extensive analysis on temperatures with the aim of evaluating changes on local extreme values (such as daily maximums) and highlighting potential relationships with other meteorological variables.

Extreme value theory (EVT) is an important topic on applied Statistics dealing with the extreme deviations from the median of probability distributions [Ghil et al., 2011]. Inherited from EVT arise several approaches applied to weather and climate extremes. This work is focused on pattern extraction of extreme data but aiming to explain steady periods of high temperatures, potentially causing heat waves. This is done through the use of quantile regression (QR) analysis. QR was introduced by Koenker and Bassett Jr [1978] as a complement to the ordinary least squares estimation (OLS). This technique was created with the aim of enhancing classical regression analysis by estimating families of conditional quantile surfaces. QR does not need to accomplish the rigorous assumptions of the conditional-mean models i.e. normality in shape, and most importantly, homogeneity of variance. QR has the additional advantage of being able to be specifically tailored to analyse non-central locations (for example daily maximums and minimums), which are precisely of main interest for extreme events analysis in both summer and winter.

Previous works on QR applied to meteorology can be found at Bjørnar Bremnes [2004] where probabilistic precipitation forecasts have been made through quantile regression methods. Friederichs and Hense [2007] applied QR to statistical downscaling for extreme precipitation events. Lee et al. [2013] successfully investigated QR models with extreme temperatures. Taillardat et al. [2016] used quantile regression forests, a generalization of random forests for quantile regression, to short-term forecasting of temperature and wind speed. The aim of the work presented here is to apply regression to separate years of summer data and to eventually combine them using a weighted average ensemble to produce a representative summer year that contains the events seen in the original data sets. Beyond developing a forecast methodology, this paper analyses the causes of temperature extreme periods during the summer time. The approach is novel in its application to the weather characterisation for building energy assessment, where as far as the authors know, quantile regression methods have not been applied yet. The knowledge gained through QR will be highly valuable in further research using building simulation and other applications of weather resilient design.

2 QUANTILE METHODS FOR SUMMER TEMPERATURES

The interest in the analysis of summer temperatures for the built environment goes beyond the calculation of demands. Extreme temperatures could cause serious overheating issues especially where high maximum temperatures during the day coincide with warmer minima at night. QR provides response to this concern by estimating the effect

of a set of weather inputs, x, on the quantiles of the temperature, y.

2.1 Quantile regression

QR considers the relationship between the input variables and the output, by performing a conditional regression in a similar way to the conditional mean function used for OLS linear regression. Similarly to OLS, the conditional median function, $Q_q(y|x)$, would be applied; where the median is the 50^{th} percentile. The expression can be straightforward extended to any quantile, q, of the empirical distribution. The quantile $q \in (0, 1)$ for y splits the data into proportions q below and $1 - q$ above: $F(y_q) = q$ and $y_q = F^{-1}(q)$. Following the parallelism, while OLS minimizes $\sum_i e_i^2$, QR minimises a sum that gives asymmetric penalties $(1 - q)|ei|$ for over-prediction and $q|ei|$ for under-prediction. The quantile regression estimator for quantile q minimises the objective function of Equation (1).

$$Q(\beta_q) = \sum_{i:y_i \geq x_i'\beta}^{N} q|y_i - x_i'\beta_q| + \sum_{i:y_i \leq x_i'\beta}^{N} (1-q)|y_i - x_i'\beta_q| \tag{1}$$

The non-differentiable function of Equation (1) requires linear programming methods for its minimization. Commonly used approaches, such as Simplex method for moderate data size or Interior Point method for larger databases, guarantee to yield a solution in a finite number of iterations. Bootstrap standard errors are often used instead of analytic standard errors, even in the case of QR errors residues are asymptotically normally distributed. Bootstrap methods are often preferable because they make no assumption about the distribution of response [Kocherginsky et al., 2012], being able to generalize QR at any case of residual distribution.

2.2 Quantile regression ensemble

Ensemble learning is a process that uses a set of models to study a common objective. All of these single models are integrated to obtain a more robust and accurate approach for temperature predictions, in addition to help to maintain a suitable uncertainty level [Mendes-Moreira et al., 2012]. The final model is generated from combining the single models or by selecting the best models in terms of accuracy [Kuncheva, 2002]. For regression problems, ensemble integration is done using a linear combination of the predictions. For QR, this is given by Equation (2),

$$Q_{Tq}(y|x) = \sum_{i=1}^{K} h_{q,i}(y|x) \cdot Q_i(y|x) \tag{2}$$

where K is the number of single QRs to make up the ensemble, q represents a specified quantile for QR, and $h_{q,i}(y|x)$ are weighted functions; $i = 1, \cdots, K$. In this case, the interest is focused on ensemble regressions with weights proportional to the distance of single QRs to the QR for the median, $Q(q50)$. Thus, $h_{q,i}(y|x)$ is given by the expression of Equation (3),

$$h_{q,i}(y|x) = \begin{cases} \alpha_i \cdot [Q_{q,i}(y|x) - Q_{50,i}(y|x)] & \text{for } q \in \text{upper QR set} \\ \alpha_i' \cdot [Q_{50,i}(y|x) - Q_{q,i}(y|x)]^{-1} & \text{for } q \in \text{lower QR set,} \end{cases} \tag{3}$$

where α_i and α_i' are normalizing coefficients such that $\sum h_i(q) = 1$. The choice of these weights aims to increase the importance of critical phases of summer temperatures that contain the highest temperatures during the day coinciding with warmer nights.

3 SUMMER YEARS ENSEMBLE: A CASE-STUDY

To evaluate the validity of the QR, a set of weather records was used. The meteorological database under study corresponds to 50 years of hourly data (1961 - 2010) collected at Heathrow weather station (London, UK) from BADC [UK Meteorological Office, 2012]. For each year, the months of April to September are selected to represent the summer period. The data available is: wind direction (wdir) in North azimuth degrees, wind speed (wspeed) in knots, cloud cover (cloud) in scale 1-8 from minimum to maximum cloud cover, air pressure (airp) in millibars, air temperature (airt) in degrees Celsius, and dew point temperature (dpt) in degrees Celsius. The air temperature distribution for each summer is summarised in Figure 1.

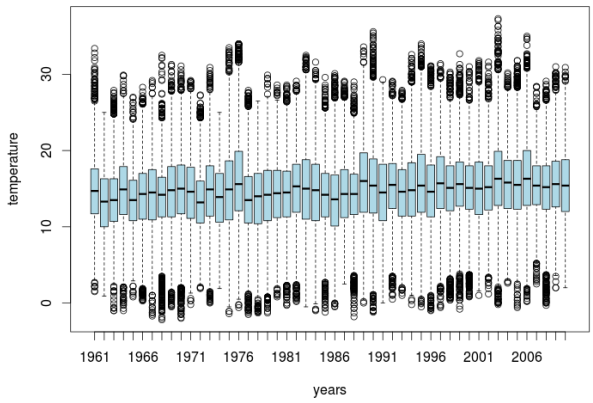


Figure 1: Comparison of temperature distribution over the 50 years under study

In the case of London (UK) the probabilistic Design Summer Years (pDSYs) [CIBSE, 2014] were proposed as a reference of warm summers. This selection was focused on different overheating metrics such as the number of hours in a building in which the temperature is above a certain threshold when occupied. Instead of mean summer dry bulb temperature, Nicol et al. [2009] suggested to use weighted cooling degree hours (WCDH) to measure the severity of warmth of a given year. The statistics for selecting pDSY based on the ascending order of WCDH [CIBSE, 2014] were based on the years from 1984 to 2006. The previous years, from 1975 to 1983, were used to determine return periods. The pDSYs given by this criterion resulted on 1989, 1976 and 2003. The year 1989 is the current CIBSE DSY representing a moderately warm year, year 1976 contains a long period of extreme summer and year 2003 contains an extreme hot event for a short period. The selection of DSY is therefore prven poor when analysing overheating in buildings [Jentsch et al., 2014].

We can check the usefulness of the QR approach by evaluating the goodness of year 1989 as a reference (current CIBSE DSY). Table 1 shows the difference between the coefficients for OLS regression and QR for quantiles 0.05, 0.5, 0.95; see Equation (1). The coefficients represented in Table 1 represents the difference in the predicted value

of temperature for each one-unit difference in the input associated with the coefficient, if the rest of inputs remain constant. This means that if the input varies by one unit, and the rest of inputs does not vary, the temperature will differ on average the quantity represented by the associated coefficient. In short, we refer to this effect represented by each coefficient as an "impact" of the input variable in the temperature.

Table 1: Coefficients of OLS and QR at 0.05, 05, and 0.95 quantiles. Year 1989.

Input	OLS	QR 0.05	QR 0.50	QR 0.95
wdir	0.0036	0.0011	0.0032	0.0038
wspeed	0.5862	0.2491	0.6100	0.6536
cloud	-0.3178	-0.1368	-0.2965	-0.4712
airp	0.1646	0.0701	0.1898	0.2201
dpt	0.8443	0.9685	0.0161	0.6825

Figure 2 shows prediction intervals for every quantile at each explanatory variable. In red is the result of the OLS regression based on the mean. We can see how the relationships in the weather database significantly change depending on the quantile, as there is no intersection between the OLS confidence interval and the QR prediction intervals.

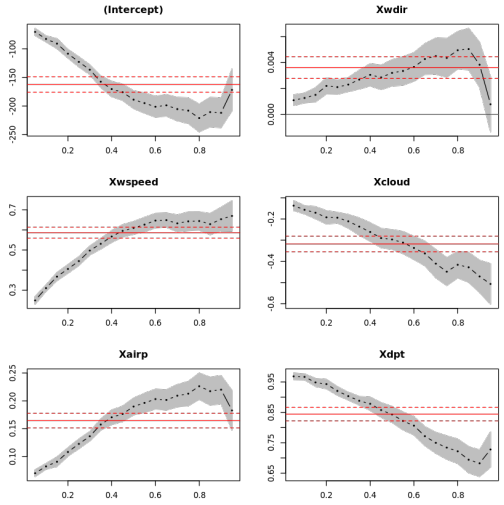


Figure 2: Impact of QR explanatory variables depending on quantile. Hourly summer temperatures, year 1989.

Wind direction does not seem to have a contribution in temperatures with a significant difference regarding OLS in most of the quantiles. The parameter values of Equation (1) for cloud cover and dew point temperature decrease when QR quantile value increases. Nevertheless, the parameter values of dew point decrease to zero with the value of quantile while the parameter values of cloud cover a negative and its decreasing implies increasing in absolute value terms. The wind speed and air pressure coefficients increase as QR has highest values. The regression scatter plot for QR and the hourly data for 1989 summer time is represented in Figure 3, where QR values for the 0.05 quantile are in red and for 0.95 are in green colour.

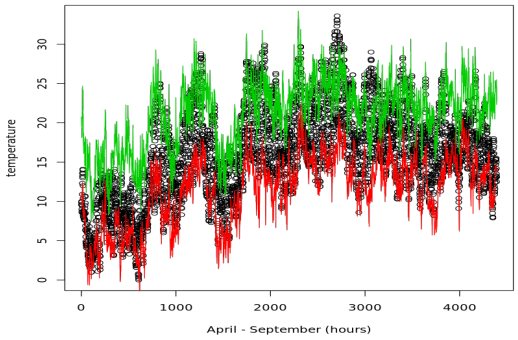
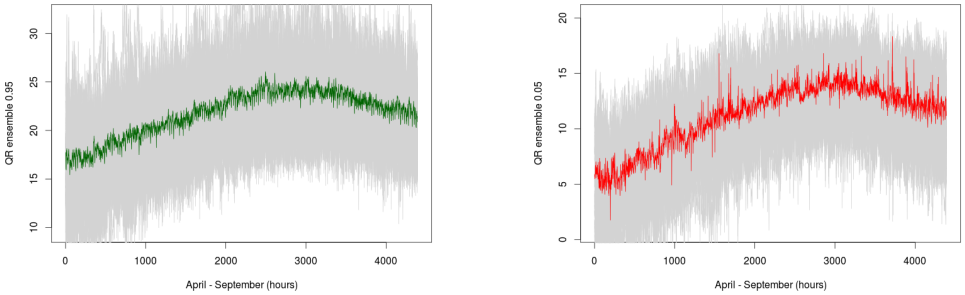


Figure 3: QR for 0.05 and 0.95 quantiles and hourly summer temperatures of 1989.



(a) Ensemble of 50 QRs at 0.95 quantile

(b) Ensemble of 50 QRs 0.05 quantile

Figure 4: Ensembles of QRs at 0.05 and 0.95 quantiles for the summer temperatures of the years 1961 - 2010

The QR outcome for one year of weather data has been shown to provide key information to better understanding the summer temperatures. The next phase of the analysis corresponds to developing a QR model ensemble for the years 1961 - 2010. As it is previously stated in equations (2) and (3) of Section 2, the ensemble weights are proportional to the distance to the median. As a result we have the two regressions, for the quantiles 0.05 and 0.95 as displayed in Figure 4. Each one is an ensemble over the predictors of 50 regression models corresponding to each of the 50 years in the database. The ensemble parameters have been tuned by cross-validation over random partitions of the data into training and test summer periods.

In addition to the natural interest on high summer temperatures, the interest goes further to also consider the higher lower QRs that can be detected by their closeness to the upper QRs. Figure 5 shows the outputs of each ensemble of Figure 4 in just one plot. This makes possible to check how the first half of July (approx. starting from hour 3,000) is when the two QRs (QR(0.05) and QR(0.95)) take highest values and also have lower distance between themselves. Consequently, heat waves are more likely during this period. Common weather patterns having an impact on the result of both ensembles are: wspeed and dpt for the QR(0.05) ensemble and wspeed, dpt, and cloud (negative relationship) for the QR(0.95) ensemble.

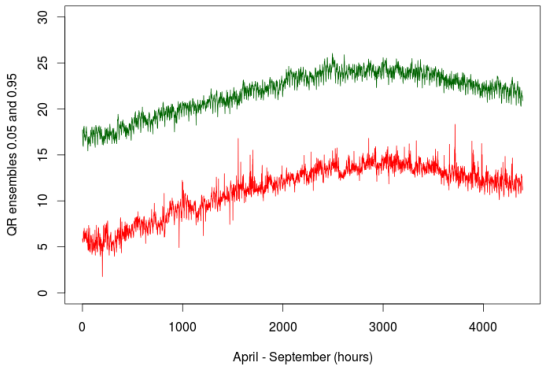


Figure 5: QR 0.05 and QR 0.95 ensembles.

4 Conclusions

The technique presented in this paper based on QR estimates rates of change for functions along or near the upper or lower boundary of the conditional distribution of temperatures. In disciplines such as building energy assessment in which the temperature swings are crucial to evaluate the severity of the weather events this seemed like a promising approach to the data processing of these series. QR models have been seen to be useful to understand the rate of changes in extreme events along with the causes of the most extreme data.

In this work, an ensemble of QR predictions based on the distance of the estimated model values with the median is also proposed. This ensemble provides a better representation of highest temperatures during the day in coincidence with warmer nights. Considering the literature of the topic, which points to a definition of heat wave that heavily emphasizes on nocturnal minimum temperatures.

The approach used here can be extended by developing parallelized QR ensembles which can be used in applications which otherwise would involve an intensive computational effort. This is the case of working with data coming from different weather scenarios and weather generators, whom seem to be widely used within the paradigm of probabilistic weather datasets.

References

Bjørnar Bremnes, J. Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review*, 132(1):338–347, 2004.

Brentan, B. M., E. Luvizotto Jr, M. Herrera, J. Izquierdo, and R. Pérez-García. Hybrid regression model for near real-time urban water demand forecasting. *Journal of Computational and Applied Mathematics*, 2016.

Christierson, B., J.-P. Vidal, and S. D. Wade. Using UKCP09 probabilistic climate information for UK water resource planning. *Journal of Hydrology*, 424:48–67, 2012.

CIBSE. Design Summer Years for London – CIBSE TM49, 2014.

Coley, D. and T. Kershaw. Changes in internal temperatures within the built environment as a response to a changing climate. *Building and environment*, 45(1):89–93, 2010.

Coley, D., T. Kershaw, and M. Eames. A comparison of structural and behavioural adaptations to future proofing buildings against higher temperatures. *Building and Environment*, 55:159–166, 2012.

de Wilde, P. and D. Coley. The implications of a changing climate for buildings. *Building and Environment*, 55:1–7, 2012.

Ebinger, J. O. *Climate impacts on energy systems: key issues for energy sector adaptation*. World Bank Publications, 2011.

Friederichs, P. and A. Hense. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review*, 135(6):2365–2378, 2007.

Ghil, M., P. Yiou, S. Hallegatte, B. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, et al. Extreme events: dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3):295–350, 2011.

IPCC. *Climate Change 2014: Mitigation of Climate Change*, volume 3. Cambridge University Press, 2015.

Jentsch, M. F., G. J. Levermore, J. B. Parkinson, and M. E. Eames. Limitations of the CIBSE design summer year approach for delivering representative near-extreme summer weather conditions. *Building Services Engineering Research and Technology*, 35 (2):155–169, 2014.

Kocherginsky, M., X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55, 2012.

Koenker, R. and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Kuncheva, L. I. Switching between selection and fusion in combining classifiers: an experiment. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 32(2):146–156, 2002.

Lee, K., H.-J. Baek, and C. Cho. Analysis of changes in extreme temperatures using quantile regression. *Asia-Pacific Journal of Atmospheric Sciences*, 49(3):313–323, 2013.

Mendes-Moreira, J., C. Soares, A. M. Jorge, and J. F. D. Sousa. Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10, 2012.

Nicol, J. F., J. Hacker, B. Spires, and H. Davies. Suggestion for new approach to overheating diagnostics. *Building Research & Information*, 37(4):348–357, 2009.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, (in press), 2016.

UK Meteorological Office. Met Office Integrated Data Archive System (Midas) land and marine surface stations data (1853-current). *NCAS British Atmospheric Data Centre*, (accessed 01/10/14), 2012.

Wilby, R., H. Orr, G. Watts, R. Battarbee, P. Berry, R. Chadd, S. Dugdale, M. Dunbar, J. Elliott, C. Extence, et al. Evidence needed to manage freshwater ecosystems in a changing climate: turning adaptation principles into practice. *Science of the Total Environment*, 408(19):4150–4164, 2010.