Data scarcity

Some food commodities do not have sufficient amount of related tweets and frequently there exist zero-tweet periods because food price is not major topic on twitter. Table 1 displays the scale of data scarcity problem on Twitter. Price measurement on these periods is impossible due to missing of data, so the estimation model should consider about how to deal with these periods. For dealing with this data scarcity problem, we decided to refresh model when there is no tweet data over several (k) days. The basic idea is to restart the model with starting price of recent average price (from n days before today) because we are not confident about the model price after a period of zero tweet data input. We tested empirical parameters as k=7 days and 2 months (n=60 days) for calculating the recent average price. Figure 1 shows how the adaptive price adjusting could work for a single case; onion price model is restarted after high peak in September 2013 due to zero-tweet period and it starts to track sharp price drop.

$$P_{t-1} = \frac{\sum_{j=t-k}^{t-1} P_j}{k} \text{ where no tweets over } n \text{ days} \quad (1)$$

Commodity	Days with zero tweet	Days with 1-2 tweet(s)	Total tweets on inflation	Total tweets on deflation
Beef	221	59	10011	5658
Chicken	207	39	3993	1594
Onion	352	72	1531	557
Chilli	312	87	975	922

Table 1. The number of days with scarce tweets out of the entire 484-day observation period (zero or 1–2 tweets mentioning food prices) and the total count of tweets quoting food prices upon inflating or deflating periods compared to the previous day

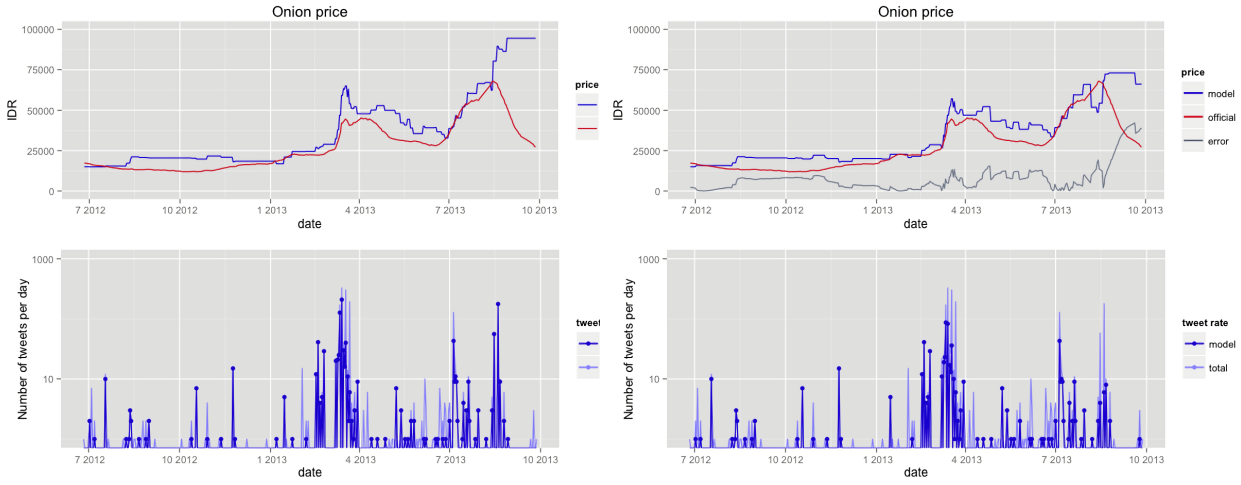

Figure 1. Data scarcity treatment: (a) Onion price modeling without tweet scarcity handling. Model price could not track a real price since near September 2013 because of rapid real price dropping in zero-tweet period. MAE = 9845.24 and r = 0.80 (p<0.001). (b) After utilizing data scarcity handling, onion price model starts to track sharp price drop after September 2013 and it shows better performance - MAE decreased to 9533.87 and r increased to 0.82 (p<0.001).