Figure S1. Identification of the optimal dataset by 10 times experiments of Lasso. 'B set (Blog set)' means a dataset including keywords and its volumes generated from web blog posts. 'C set (Twitter set)' means a dataset including keywords and its volumes generated from twitter. 'A set' has been constructed as average of keyword volumes generated from web blog posts and twitter, respectively. 'D, E, F and G set' are merged dataset based on A, B and C set.

MSE, mean squared error
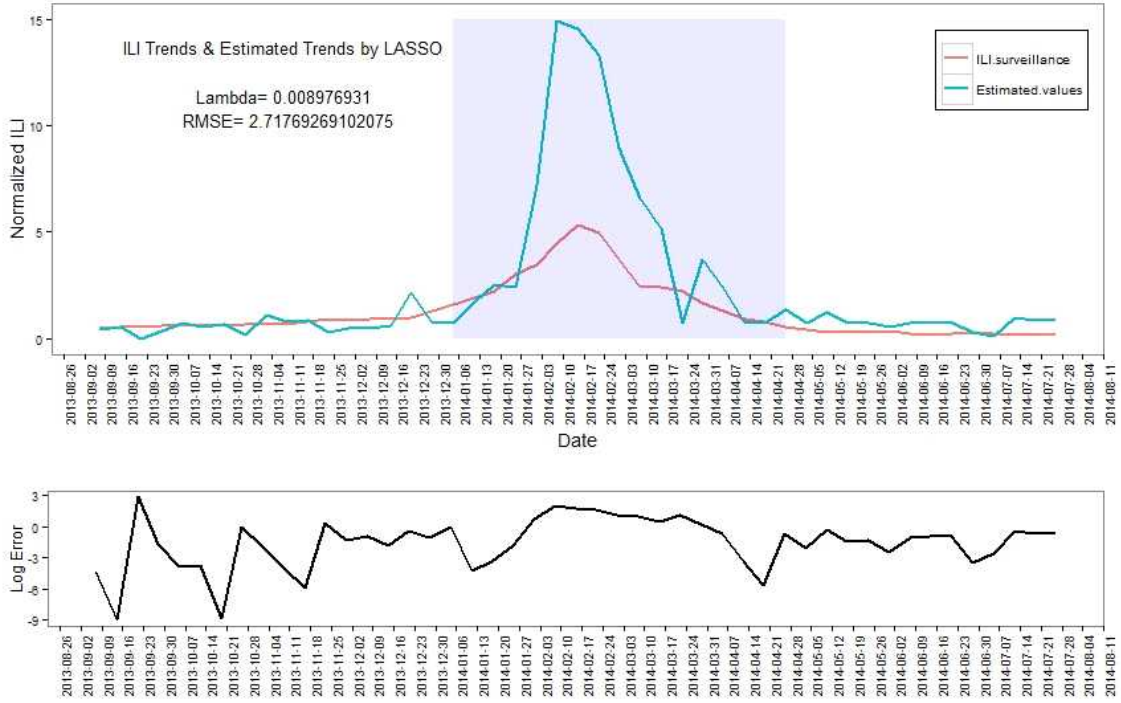
A Set : (Blog + Twitter)/2

B Set : (Blog)

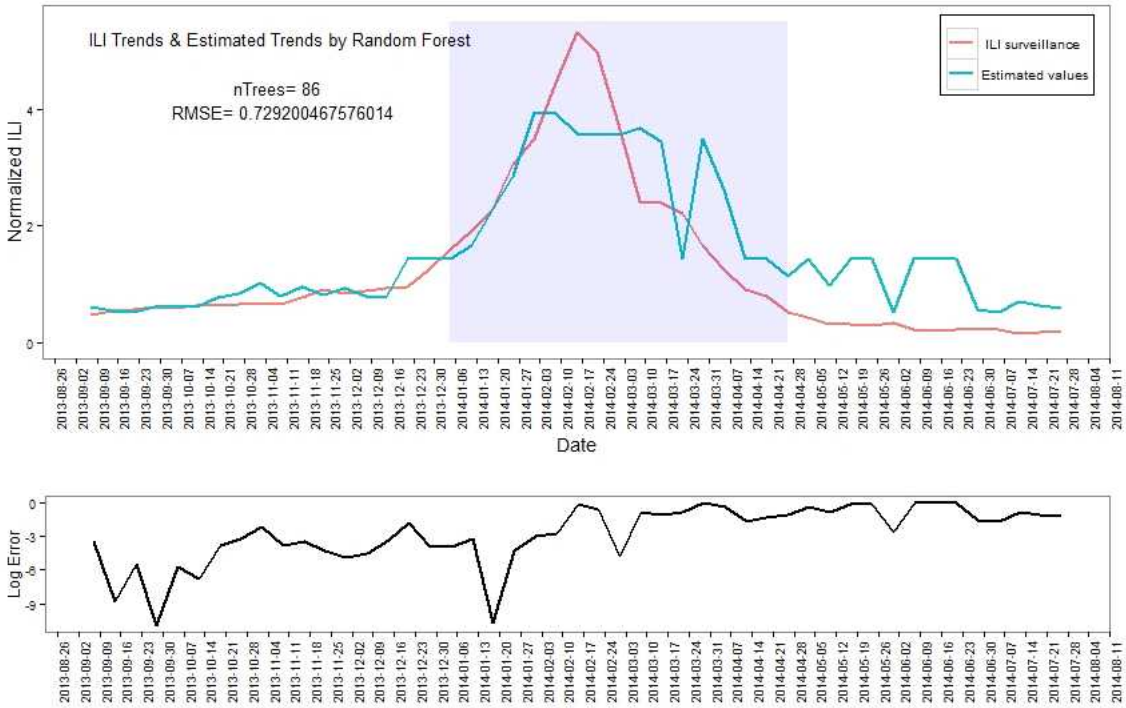C Set : (Twitter)

D Set : A Set & B Set

E Set : A Set & C Set

F Set : B Set & C Set

G Set : A Set & B Set & C Set

Figure S2. Detection and error based on the lasso model.

Figure S3. Detection and error based on the Random Forest model.