Does Weather Matter? Causal Analysis of TV Logs

Shi Zong† Branislav Kveton‡ Shlomo Berkovsky§ Azin Ashkan† Nikos Vlassis† Zheng Wen†

†The Ohio State University, Columbus, OH, USA ‡Adobe Research, San Jose, CA, USA §CSIRO, Epping, NSW, Australia ‡Google Inc., Mountain View, CA, USA
zong.56@osu.edu, {kveton, vlassis, zwen}@adobe.com, Shlomo.Berkovsky@data61.csiro.au, azin@google.com

ABSTRACT
Weather affects our mood and behaviors, and many aspects of our life. When it is sunny, most people become happier; but when it rains, some people get depressed. Despite this evidence and the abundance of data, weather has mostly been overlooked in the machine learning and data science research. This work presents a causal analysis of how weather affects TV watching patterns. We show that some weather attributes, such as pressure and precipitation, cause major changes in TV watching patterns. To the best of our knowledge, this is the first large-scale causal study of the impact of weather on TV watching patterns.

1. INTRODUCTION
Weather affects our mood, and thus human behaviors. One of the pronounced examples is the seasonal affective disorder – prolonged lack of sunlight that can depress people [2]. Weather indirectly affects various aspects of our lives: work and study, pur-posed several practical ways for estimating the causality of weather on TV watching behavior, and observe high correspondence between our findings.

2. CAUSAL ANALYSIS
The problem of estimating causal effects from observational data is central to numerous disciplines [3]. It can be formalized as follows. Let \(\{1, \ldots, n\}\) be a set of \(n\) units \(i\), such as individuals. Let \(T_i \in \{0, 1\}\) indicate the treatment of unit \(i\). That is, \(T_i = 0\) if unit \(i\) is control and \(T_i = 1\) if the unit is treated. Then unit \(i\) has two potential outcomes, \(Y_i(1)\) if the unit is treated and \(Y_i(0)\) otherwise. The unit-level causal effect of the treatment is the difference in potential outcomes, \(\tau_i = Y_i(1) - Y_i(0)\), and the average treatment effect on treated (ATT) is \(E[Y_i(1) - Y_i(0)]\) across treated units. That is, if \(E[Y_i(1) - Y_i(0)]\) cannot be directly computed, because \(Y_i(0)\) is unobserved in treated units \(\{i : T_i = 1\}\).

Since the assignment to treatment and control groups is usually not random, \(E[Y_i(1) - Y_i(0)]\) is a poor estimate of \(E[Y_i(1) - Y_i(0)]\). A key challenge in causal analysis is to eliminate the resulting imbalance between the distributions of treated and control units. A popular approach to balancing the two distributions is nearest-neighbor matching (NNM) [4]. In this work, we match each treated unit to its nearest control unit based on their covariates, and then the response of the matched unit serves as a counterfactual for the treated unit. In particular, the ATT is estimated as

\[
\text{ATT} \approx \frac{1}{n_T} \sum_{i=1}^{n_T} (Y_i(1) - Y_{\pi(i)}(0)),
\]

where \(n_T = \sum_{i=1}^{n_T} T_i\) is the number of treated units, \(Y_i(1)\) is the observed response of treated unit \(i\), and \(Y_{\pi(i)}(0)\) is the observed response of the matched control unit \(\pi(i)\). The covariate of unit \(i\), \(X_i\), should be chosen such that its potential outcomes are statistically independent of \(T_i\). In this case, the estimate in (1) resembles that of a randomized experiment.

3. EXPERIMENTS
In this work, we use a dataset gathered by an Australian national IPTV provider. We obtained the complete Australia-wide logs for a period of 26 weeks, from February to September 2012 [5]. The dataset contains 10M viewing events of about 40k users, who watched more than 11k unique programs in 14 TV genres. We also obtained the geographic locations of the users and matched them with their weather. We further extracted eight attributes that characterize various aspects of weather (Table 1).

3.1 Experimental Setup
We apply the causality framework from Section 2 to our dataset. As we explain our setup, we illustrate it on an example query “does high temperature cause watching more Dramas?”. The units \(i\) are TV watching events and we estimate the causal effect of weather at these events. The treatment \(T_i\) indicates the treatment weather at event \(i\). In our example, \(T_i = \{\text{high temperature at event } i\}\). We define one treatment variable for each weather attribute, and get eight weather-attribute treatments. For each treatment, we treat the events in the tail of the distribution of that attribute. That is, if the tail of the distribution is on the left (right), we assign 20% of the events with the lowest (highest) values to the treatment group. We
### Table 1: Our weather attributes with their treatment groups.

<table>
<thead>
<tr>
<th>Weather attribute</th>
<th>Treated</th>
<th>Weather attribute</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>High (H)</td>
<td>Pressure</td>
<td>Low (L)</td>
</tr>
<tr>
<td>Feels-like temper</td>
<td>High (H)</td>
<td>Humidity</td>
<td>Low (L)</td>
</tr>
<tr>
<td>Wind speed</td>
<td>High (H)</td>
<td>Visibility</td>
<td>Low (L)</td>
</tr>
<tr>
<td>Cloud cover</td>
<td>High (H)</td>
<td>Precipitation</td>
<td>High (H)</td>
</tr>
</tbody>
</table>

Figure 1: (a) Normalized ATTs of pressure on 8 most popular genres with TV-genre (blue) and latent (red) user profiles. (b) Significant changes in normalized ATTs of all weather treatments on 8 most popular genres. Significant increases are red, significant decreases are blue, and insignificant effects are gray. We report results for both TV genre (left) and latent (right) user profiles.