Get to the Bottom: Causal Analysis for User Modeling

Shi Zong
The Ohio State University
Columbus, OH, USA
zong.56@osu.edu

Branislav Kveton
Adobe Research
San Jose, CA, USA
kveton@adobe.com

Shlomo Berkovsky
CSIRO
Eveleigh, NSW, Australia
Shlomo.Berkovsky@csiro.au

Azin Ashkan*
Google Inc.
Mountain View, CA, USA
azin@google.com

Zheng Wen
Adobe Research
San Jose, CA, USA
zwen@adobe.com

ABSTRACT

Weather affects our mood and behavior, and through them, many aspects of our life. When it is sunny, people become happier and smile, but when it rains, some get depressed. Despite this evidence and the abundance of weather data, weather has mostly been overlooked in the machine learning and data science research.This work shows how causal analysis can be applied to discover the effects of weather on TV watching patterns and how it can be applied for user modeling. We make several contributions. First, we show that some weather attributes, e.g., pressure and precipitation, cause significant changes in TV watching patterns. Second, we compare the results obtained for different levels of user granularity and different types of users.Th is showcases that causal analysis can be a valuable tool in user modeling. To the best of our knowledge, this is thefi rst large-scale causal study of the impact of weather on TV watching patterns.

KEYWORDS

Weather; Causal Analysis; User Modeling

1 INTRODUCTION

Weather affects our mood and, thus, human behavior. One of the pronounced examples is the *seasonal affective disorder* – prolonged lack of sunlight that can potentially depress people [21]. Although indirectly, weather affects various aspects of our lives, including our work and study, purchasing behavior, and more. For example, Hirshleifer [8] and Saunders [26] found that stock returns on cloudy days are lower than on sunny days. Similarly, Murray *et al.* [19] and Parsons [20] discovered strong effects of sunlight on consumer expenditure. Social research also linked weather conditions to crime [7] and even to suicide rates [3].

In this work, we extend our early work [29] and set out to thoroughly examine the effects of weather on the TV watching activity. It is reasonable to assume that the overall TV watching levels are weather dependent, e.g., people watch more TV when it rains, and this has been partially shown by Barnett *et al.* [4] and Roe and Vandebosch [23]. However, our main goal is to further assess the effect of weather on the TV watching patterns.Th at is, we are rather interested to explore whether people watch *different genres* of programs in different weather conditions.

We posit that this knowledge can have several important implications. First, marketers may be willing to adjust the content and ratio of the advertisements to the target audience to whom they will be exposed [10]. Second, the technical configuration of the communication network, e.g., replica placement on content-delivery network servers, may be tuned to facilitate a more efficient content delivery [9].Th ird, TV content and video-on-demand service providers may benefit from this knowledge and adapt their recommendations accordingly [28].

Several prior works incorporated weather into personalized systems [1, 2, 5]. We would like to stress three limitations of these works. First, the weather was treated as an auxiliary *contextual* dimension rather than a parameter that directly impacts user behavior. Second, relatively simple correlation-based methods were applied to discover the impact of weather on user behavior.Th ird, prior works used only small-scale datasets collected exclusively for research purposes. In fact, the largest weather-related dataset, previously used in personalization research, contains fewer than 5k ratings [5]. We believe that it is pivotal to thoroughly examine the factors affecting user behavior, as these are likely to improve the quality of the personalization. We also believe that causal analysis and large-scale data are the necessary tools for such an examination; these may uncover hidden biases in real-world problems.

To this end, we advocate in this work the use of causal analysis for modeling the impact of weather on observable TV watching behavior. To the best of our knowledge, this is thefi rst work to apply causal analysis to a nation-scale dataset containing more than *10M watching events* of more than *0.6M users*. We propose and apply an efficient technique for learning the causal dependencies between weather conditions and the users' TV watching behavior patterns. We also discuss about the motivation for applying causal analysis. Hence, the main contribution of our work lies in demonstrating and validating the application of causal analysis for the purposes of modeling dependencies in user behavior.

*Work done while at Technicolor Research.

UMAP'17, July 9–12, 2017, Bratislava, Slovakia

DOI: http://dx.doi.org/10.1145/3079628.3079688

2 MATCHING FOR CAUSALITY

The problem of estimating causal effects from observational data, such as changes due to weather conditions, is central to many disciplines [18, 22, 27]. 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.Th at is, $T_i = 0$ if unit i is *control* and $T_i = 1$ if the unit is *treated.*Th en unit i has two *potential outcomes*, $Y_i(1)$ if the unit is treated and $Y_i(0)$ otherwise.Th e 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

$$\mathbb{E}_{i:T_i=1}[\tau_i] = \mathbb{E}_{i:T_i=1}[Y_i(1)] - \mathbb{E}_{i:T_i=1}[Y_i(0)]\,,$$

where $\mathbb{E}_{i:T_i=1}[Y_i(1)]$ is the expected outcome of treatment on treated units and $\mathbb{E}_{i:T_i=1}[Y_i(0)]$ is the expected outcome of not being treated on treated units. Note that $\mathbb{E}_{i:T_i=1}[\tau_i]$ 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, the expected outcome of not being treated on control units, $\mathbb{E}_{i:T_i=0}[Y_i(0)]$, is a poor estimate of the expected outcome of not being treated on treated units, $\mathbb{E}_{i:T_i=1}[Y_i(0)]$.Th e key challenge in causal analysis is to eliminate the resulting imbalance between the distributions of treated and control units. A popular approach of balancing the two distributions is the *nearest-neighbor matching (NNM)* [6, 13, 16, 24]. In this work, we match each treated unit to its nearest control unit based on its covariates, and then the response of the matched unit serves as a *counterfactual* for the treated unit. In particular, the ATT is estimated as

$$\text{ATT} \approx \frac{1}{n_T} \sum_{i:T_i=1} \left(Y_i(1) - Y_{\pi(i)}(0) \right)\,, \tag{1}$$

where $n_T = \sum_{i=1}^{n} T_i$ 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, which is essentially a d-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^d$, should be chosen such that the potential outcomes of unit i are statistically independent of T_i given $\mathbf{x}_i$. In this case, the estimate in Eq. (1) resembles that of a randomized experiment.

3 CAUSAL EFFECTS OF WEATHER ON TV CONTENT

In this section, we discuss how to apply the NNM framework from Section 2 to analyze the causal effects of weather on TV watching. We illustrate the methodology with an example query "*does high temperature cause watching more drama?*". In Section 3.1, we introduce the notions of treatment, control, and potential outcomes. We justify our choice of covariates in Section 3.2 and explain our NNM method in Section 3.3.

3.1 Treatment, Control, and Outcomes

Our *units* (events of interest) are TV watching events i; and we are interested in estimating the causal effect of weather on these events. The *treatment* T_i is an indicator of the treatment weather at event i, such as that the temperature is high. We denote the *potential outcomes* at event i under control and treatment by $Y_i(0)$ and $Y_i(1)$, respectively.Th e potential outcomes are indicators of the watched content under control and treatment. For instance, in our example query, the treatment and potential outcomes are

$$T_i = \mathbb{1}\{\text{temperature is high at event } i\}\,,$$
$$Y_i(0) = \mathbb{1}\{\text{drama watched at event } i \text{ if the temperature is low}\}\,,$$
$$Y_i(1) = \mathbb{1}\{\text{drama watched at event } i \text{ if the temperature is high}\}\,.$$

We discuss how to determine high and low temperatures in Section 5.1.

We measure the *effect* of treatments by the ATT in Eq. (1). In our domain, the ATT is the expected increase in the frequency of watching some content due to the treatment weather, such as the expected increase in the frequency of watching drama due to high temperature. If the ATT is significantly above zero, we claim that high temperature *increases* the frequency of watching drama. If the ATT is significantly below zero, we claim that high temperature *decreases* the frequency of watching drama. Finally, if the ATT is near zero, we claim that high temperature has *no effect* on watching drama. We provide a detailed discussion about how to measure the significance of effects in Section 5.2.

3.2 Covariates and Ignorability

A key step in causal analysis is to break the dependence between potential outcomes and treatments, in order to mimic a randomized experiment.Th is can be done under the assumption of *unconfoundness* or *ignorability* [22, 25].Th e ignorability assumption says that the potential outcomes are statistically independent of the treatment given the covariates. In particular,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid \mathbf{x}_i$$

for any TV watching event i, where $\mathbf{x}_i \in \mathbb{R}^d$ are the covariates of event i. In this paper, the covariates are the profile of the user at event i, the location of event i, and the time of event i.The profile is the distribution over watched TV genres of the user (see Section 5.1), which clearly affects $(Y_i(0), Y_i(1))$.Th e time affects the availability of content. For instance, as shown in Fig. 1, the frequency of watching TV genres can change dramatically over time. Finally, both the time and location are good predictors of the weather, which is the treatment.

Our ignorability assumption says that what the user would have watched under different weather conditions at event i, $(Y_i(0), Y_i(1))$, depends on the profile of the user at event i, the location of event i, and the time of event i; but does not depend on treatment T_i. For example, the fact that the temperature is high should not correlate with what the user would have watched under high and low temperatures.Th e ignorability assumption is hard to validate in practice [25], but we believe that our choice of covariates renders it reasonable in our setting. Properly chosen covariates reduce, if not entirely eliminate, the dependence between potential outcomes and the treatment.Th en the observed effect is likely due to the causation through the treatment.

We illustrate our ignorability assumption on an example. Consider a family where the parents like to watch drama. When the temperature is high, the children play outside and the parents watch drama. When the temperature is low, the children are at home and the whole family watches family movies. In this case, the potential

Algorithm 1 Large-scale matching of treatment and control events using random partitioning.

// Random partitioning
Partition covariates $\{\mathbf{x}_i\}_{i=1}^n$ into k random clusters $\{C_\ell\}_{\ell=1}^k$ such that $n/k \leq |C_\ell| \leq n/k + 1$ for all $\ell \in [k]$

// Matching
for all $\ell = 1, \ldots, k$ **do**
  while $|C_\ell \cap \{i : T_i = 1\}| > 0$ **do**
    Choose a random treated event i from
      $C_\ell \cap \{i : T_i = 1\}$
    Find the nearest control event j to event i in
      $C_\ell \cap \{i : T_i = 0\}$
    Match the events i and j, $\pi(i) \leftarrow j$
    Remove the events from C_ℓ, $C_\ell \leftarrow C_\ell \setminus \{i, j\}$
  end while
end for

outcomes are determined solely by the profile of the family. In particular, they are determined independently of whether event i is treated or not, implying that the family subconsciously decides on what to watch when the temperature is high or low before either of these happens.Th erefore, the potential outcomes are statistically independent of the instance of the weather at event i and our ignorability assumption holds.

3.3 Efficient Matching via Clustering

The number of TV watching events can be large, on the order of millions. Naive implementations of NNM from Section 2 are impractical in this setting because their running time is $O(n^2)$. In our work, we propose a computationally-efficient variant of NNM based on the idea of *quantization* [11].

Our algorithm for NNM is presented in Algorithm 1.Th e algorithm has two main stages. First, we randomly partition all covariates into k clusters $\{C_\ell\}_{\ell=1}^k$. Second, we match randomly chosen treated events in each cluster C_ℓ to their nearest control events in that cluster, until no treated events are left. We choose treated events randomly to avoid biases in the matching due to a particular order.Th e clusters are also chosen randomly. When the number of events n is large and the number of clusters is reasonably small, we expect the distribution of the covariates in each cluster to closely resemble that of $\{\mathbf{x}_i\}_{i=1}^n$, and therefore NNM on the clusters should be similar to that on $\{\mathbf{x}_i\}_{i=1}^n$.

Our matching algorithm is surprisingly simple, effective, and its computational cost is only $O\left(k(n/k)^2\right) = O\left(n^2/k\right)$.Th e number of clusters k can be used to trade off the computational complexity of NNM for its quality. When k is chosen appropriately, such as $k = \sqrt{n}$, the computation cost is $O\left(n^{3/2}\right)$. In our experiments, $n \approx 10^7$ and we choose $k = 200$.Th is allows us tofi nd nearest neighbors for millions of treated events in less than an hour, on a computer with 16 GB main memory and 2.5 GHz Intel Core i7 processor. We experimented with other values of k and our causal findings are not sensitive to the choice of k, only the computational cost of the matching is. We investigate the quality of the proposed algorithm in Section 5.

4 DATASET

In this work, we used a dataset gathered by a leading Australian national TV broadcaster.Th e broadcaster offers two services: live broadcast, and a catch-up TV service available through a Web-based portal, which allows users to watch on-demand any program that they may have missed. We obtained the complete Australia-wide portal logs for a period of 26 weeks, from February to September, 2012.Th e original dataset includes more than 21.4M viewing events of about 1.3M unique users, who collectively watched more than 11k unique programs. We randomly choose 50% of these users and conduct all experiments on this set of users.

Each viewing event is represented by the user's IP address, viewing date (with no time stamp), and program ID. We use the IP addresses of the users as their unique identifiers, since little user information is available[1]. In addition to the program ID, program meta-data contains its title, duration, publication date, and a single genre tag from 14 genres (see Table 1).Th ese genre tags are labeled by the broadcaster. No information about the watched portion of the program is available. A more detailed description of the dataset can be found in Xu *et al.* [28].

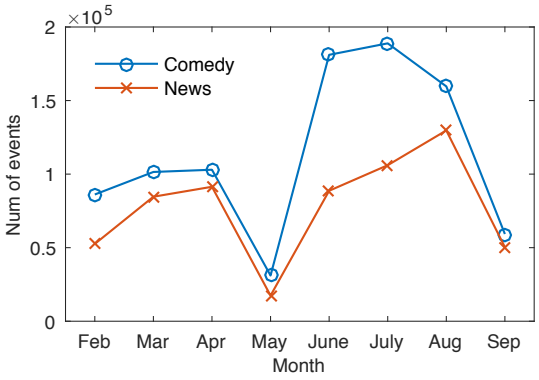


Figure 1: Number of events for *Comedy* and *News* over months in our dataset.

Table 1 shows the distribution of viewing events across TV genres. Not surprisingly, some genres are more popular than others. We also observe in Fig. 1 that the relative frequency of TV genres changes over time[2]. In April, the number of watched *Comedies* is almost the same as that of *News*. In June, the number of *Comedies* is twice as high as that of *News*.Th ese imbalances provide us with an opportunity to analyze what factors lead to the changes of users' watching behavior.

To collect the required weather data, we used the IP2Location[3] API to convert the IP addresses into the geographic locations of the users: longitude, latitude, city, and state. Given the locations and the dates of the viewing events, we used the WorldWeatherOnline[4] API to obtain the weather logs for the dates of the events.

[1] Such identifiers may be noisy, as multiple users (e.g., household members) may share the same IP and one user (e.g., at home and at work) may have multiple IPs. However, these are real data gathered by the broadcaster.
[2] The drop in May is due to a technical problem that caused some data loss.
[3] http://www.ip2location.com
[4] http://www.worldweatheronline.com

Weather conditions in Australia change quite dramatically across different geographic locations.Th e weather inland is hotter and drier than along the coast.Th e northern regions that are closer to the equator are warmer than the southern ones. To get a better idea of how the weather in Australia looks like, we anecdotally map weather conditions in Australia to other locations all over the world. We observe that the weather in the inner part of Australia is comparable to Sahara and Sonora deserts.Th e weather in the regions near the coast is similar to Southern California, San Francisco Bay Area, and Florida in the United States. We describe weather conditions by *weather attributes*.Th e weather attributes are eight numerical features extracted from weather logs that characterize different aspects of weather, such as *Temperature*, *Pressure*, and *Humidity* (see Table 1).

5 CAUSAL MODELING OF WEATHER EFFECT

Causal analysis is important for user modeling, as it can help determine factors that drive changes in user behavior. Liang *et al.* [17] showed that the idea of causal analysis can be used to model user exposure in recommender systems. In this section, we estimate the effect of weather on users' TV watching behavior by conducting causal analysis of weather attributes from Section 4. In Section 5.1, we describe our experimental setup. In Section 5.2, we analyze the average causal effects on the data of the whole of Australia. In Section 5.3, we conduct causal analysis at the level of individual users; and in Section 5.4, we conduct causal analysis on two groups of users. Finally, in Section 5.5, we showcase the necessity for causal inference.

5.1 Experimental Setup

We define one treatment variable for each weather attribute in Table 1.Th erefore, we have 8 weather-attribute treatments. In each attribute, we treat the events in the tail of the distribution of that attribute. Specifically, if the tail of the distribution is on the low (high) end of the range, we consider the 20% of the events with the lowest (highest) values of the attribute as the treatment group and the rest as the control group (see Fig. 2). We denote these high-value and low-value treatment groups by "H" and "L", respectively.The position of the tail is estimated automatically from the skewness of the distribution. We list all treatments in Table 1.

The events in the tail of the distribution are extreme by definition, and therefore they are a natural candidate for being chosen as treatments. We choose the 20% cutoffs separately for each month, as the boundaries of the "high" and "low" weather attribute values change over time, e.g., summer vs. winter temperature.Th e 20% cutoff is chosen such that the number of treated events is reasonably large. We experimented with cutoffs of 15% and 25%, and the results were similar to those of the chosen 20% cutoff.

We define a pair of potential outcomes $(Y_i(0), Y_i(1))$ for each TV genre in Table 1.Th erefore, we estimate 14 effects. In this setting, the ATT in Eq. (1) is the expected change in the frequency of watching a given TV genre due to being treated. For example, consider the effect of *high temperature* weather on watching *Dramas* that was discussed in Section 3.1. Note that we estimate the effect on TV program genres, rather than on individual TV programs, as many programs may not have enough treated events to allow accurate causal analysis.Th erefore, we experiment with a higher genre-level of content granularity.

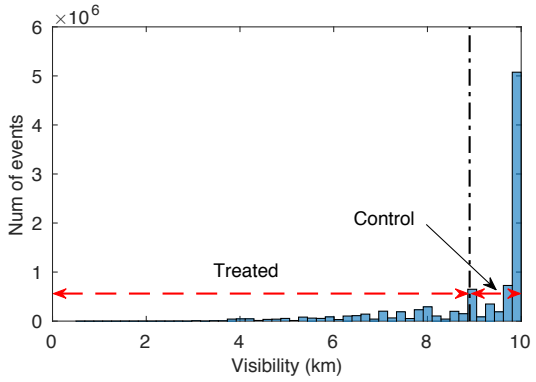


Figure 2:Th e treatment and control groups for weather attribute *Visibility*.Th e treated events are 20% of all events with lowest visibility.Th e control events are the remaining 80% of all events.

Finally, as discussed in Section 3.2, the covariates of a watching event i are the profile of the user at i, the time of i, and the location of i. In this paper, the user profile is a vector of the frequency of watching TV genres.Th is profile is a 14-dimensional vector $(m_{u,1}/m_u, \ldots, m_{u,14}/m_u)$, where $m_{u,y}$ is the number of events where user u watches genre y and $m_u = \sum_{i=1}^{14} m_{u,i}$ is the total number of events of u. Such a profile naturally captures high-level preferences of the user[5].

5.2 Causal Analysis on the Population of Whole Australia

In ourfi rst experiment, we conduct causal analysis on the whole population of Australia.Th at is, for every treatment j and genre y in Table 1, we match all treated events to control events by Algorithm 1 and then estimate the ATT in Eq. (1). We denote the resulting ATT by $\text{ATT}_{j,y}$ and refer to it as the *empirical effect of treatment j on genre y*.

Note that $\text{ATT}_{j,y}$ is random, because the matching π computed by Algorithm 1 is random.Th us, we need to be careful when we evaluate the estimated effect. Consider the following example. Suppose that only one event is treated, where the user does not watch drama; and that this event is randomly matched to another event, where the user watches drama.Th en it may seem that the treatment leads to watching no drama. While this may be true, it is unlikely because this effect is estimated from only one matched pair of treated and control events. Below we propose a variant of $\text{ATT}_{j,y}$ that allows us to eliminate statistically insignificant effects.

[5]We also experimented with another type of a user profile, which was estimated using SVD from the watched TV programs.Th is profile is inspired by the low-dimensional representation of latent user profiles in matrix factorization [14]. We observe similar patterns to those in this paper, and therefore do not report them.

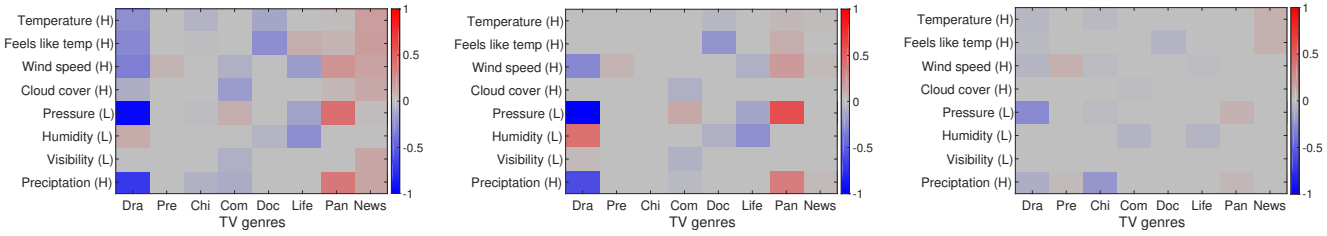
Category	Frequency	Category	Frequency
Drama	19.51%	Pre-school	19.31%
Children	17.01%	Comedy	11.37%
Docs	10.61%	Lifestyle	8.06%
Panel	5.95%	News	4.10%
Arts	2.69%	Education	0.58%
Kids	0.50%	Sport	0.24%
Indigenous	0.05%	Shop	0.02%

TV genres.

Weather attribute	Treated
Temperature	High
Feels-like temperature	High
Wind speed	High
Cloud cover	High
Pressure	Low
Humidity	Low
Visibility	Low
Precipitation	High

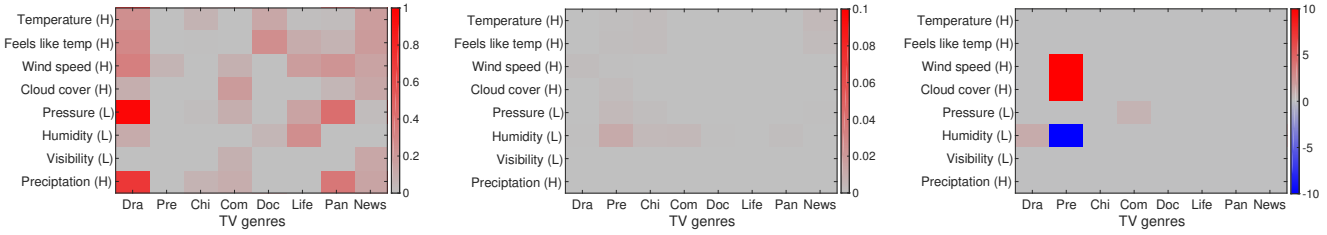
Weather attributes and their treated values.

Table 1: TV genres and weather attributes as described in Section 4.



(a) Whole of Australia data group (~2M treated events) (b) *Doctor Who* group (~0.7M treated events) (c) *Peppa Pig* group (~0.5M treated events)

Figure 3: High-probability effects $\widetilde{\mathrm{ATT}}_{j,y}$ in 8 most popular genres due to 8 weather-attribute treatments for: (a) whole of Australia, (b) users who watched *Doctor Who*, and (c) users who watched *Peppa Pig*.The effects are multiplied by 100 and can be interpreted as changes in the percentage of watching TV genres. "H" and "L" denote high and low value treatments in Table 1.



(a) Whole of Australia as a group (total effect: 7.767) (b) Individual users as subgroups (total effect: 0.045) (c) User No. 15051 (~200 treated events)

Figure 4: (a, b) Effects in context G, $\widetilde{\mathrm{ATT}}^{G}_{j,y}$, when G is the whole population of Australia and individual users, respectively. (c) High-probability effects $\widetilde{\mathrm{ATT}}_{j,y}$ for user No. 15051.The effects are multiplied by 100 and can be interpreted as changes in the percentage of watching TV genres.

To eliminate statistically insignificant effects, we propose *high-probability effect of treatment j on genre y*,

$$\widetilde{\mathrm{ATT}}_{j,y} = \begin{cases} \max\{\mathrm{ATT}_{j,y} - c \cdot \mathrm{se}_{j,y}, 0\}, & \mathrm{ATT}_{j,y} > 0; \\ \min\{\mathrm{ATT}_{j,y} + c \cdot \mathrm{se}_{j,y}, 0\}, & \mathrm{ATT}_{j,y} < 0, \end{cases} \quad (2)$$

where $\mathrm{se}_{j,y}$ is the standard error in the estimate of $\mathrm{ATT}_{j,y}$, and $c > 0$ is a tunable parameter that controls the degree of confidence. This metric can be justified as follows. If the estimated effect is positive, $\mathrm{ATT}_{j,y} > 0$, and significant in the sense that it is larger than c times the standard error, then this effect should be reported as $\widetilde{\mathrm{ATT}}_{j,y} > 0$. Similarly, if the estimated effect is negative, $\mathrm{ATT}_{j,y} < 0$, and significant in the sense that it is smaller than c times the standard error, then this effect should be reported as $\widetilde{\mathrm{ATT}}_{j,y} < 0$. In all other case, $\widetilde{\mathrm{ATT}}_{j,y} = 0$ and the effect is not significant.

In our experiment, we choose $c = 4$. From the central limit theorem, $\mathrm{ATT}_{j,y}$ is close to normally distributed when the number

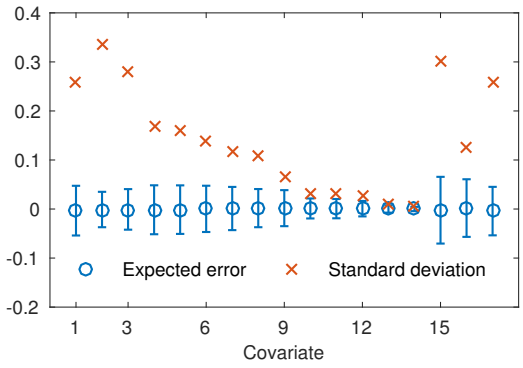


Figure 5: Comparison of the expected errors in matching in each covariate and the standard deviation of that covariate.

of treated events is large, at least 30 [15]. In this case, $\text{ATT}_{j,y} - 4 \cdot \text{se}_{j,y}$ can be viewed as a high-probability lower bound on the true effect if this effect is positive, and $\text{ATT}_{j,y} + 4 \cdot \text{se}_{j,y}$ can be viewed as a high-probability upper bound on the true effect if this effect is negative.Th ese upper and lower bounds hold with probability of at least $1 - 10^{-4}$. When the number of treated events is small, less than 30, we set $\widetilde{\text{ATT}}_{j,y} = 0$. In this case, it is unreliable to substitute the unknown standard deviation of $\text{ATT}_{j,y}$ with its empirical estimate $\text{se}_{j,y}$ and guarantees cannot be provided in general [15], unless we make strong assumptions on the distribution of $Y_i(1) - Y_{\pi(i)}(0)$ in Eq. (1).

We report $\widetilde{\text{ATT}}_{j,y}$ for all treatments j and genres y in Fig. 3(a). Note that at this granularity level, there are around 2M treated events for each weather attribute. We observe some insightful trends. For example, there is around 1% decrease in watching *Dramas* when the pressure is low and precipitation is high. A reason for this may be that rainy days tend to make people sad and they prefer not to watch dramas, which are unlikely to cheer them up. Although a 1% decrease may seem small, considering the fact that *Dramas* account for about 20% of all watching events, this is about a 5% relative decrease in this genre.Th is decrease observed on rainy days comes at the account of the more entertaining *Panel* programs, which increase by about 0.6%. Considering that *Panel* only account for 6% of the watching events, their relative increase is as high as 10%. We were not able to validate these intuitive explanations. Nevertheless, they suggest that our weather attributes may be indicative of complex human behavior patterns, which would be difficult to uncover otherwise.

All matching methods should be followed by an assessment of their quality (Section 2). In Fig. 5, we report the quality of our matching in each covariate. In particular, we report the expected error in matching in each covariate and compare it to the standard deviation of that covariate. We observe two trends. First, all expected errors are close to 0, which means that we do not introduce systematic biases in any covariate. Second, standard deviations of all errors are consistently smaller than that of the corresponding covariate.Th is shows that we match accurately in all covariates. Hence, our estimated effects are likely to be causal.

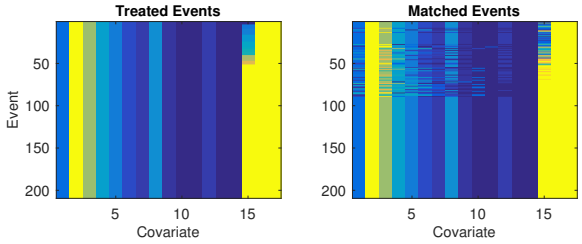


Figure 6: Treated covariates of user No. 15051 with the covariates of their matched events.

5.3 Causal Analysis on Individual Users

One of the goals of user modeling is to inform personalization. In this subsection, we partition our dataset into individual users and conduct causal analysis of these users. From the technical point of view, the causal analysis of a single user is no different from that in Section 5.2.Th e only difference is that the set of treated events in the estimate of ATT in Eq. (1) includes only the events of that user. We denote the resulting high-probability effect in Eq. (2) by $\widetilde{\text{ATT}}^k_{j,y}$ and refer to it as the *high-probability effect of treatment j on genre y of user k*.

We note that several users in our dataset have very significant and different effects from that in Fig. 3(a), which reports high-probability effects for whole of Australia. In Fig. 4(c), we report the high-probability effects on an individual with more than 2,000 watching events. Even for this user, there are only around 200 treated events for each weather attribute. Unfortunately, most users in our dataset have much less watching events than 2,000. The number of users with more than 2,000 events is only 41, among 578,308 users. If the user does not have enough treated events, the high-probability effect in Eq. (2) is likely to be zero, based on the discussion in Section 5.2. In Fig. 3(a), we do not observe significant effects for *Pre-school* genre when weather changes. However, in Fig. 4(c), high cloud cover, high wind speed, and low humidity all have large affects on the user's preference towards *Pre-school* TV programs.Th is indicates that conducting causal analysis on individual users can help reveal more complicated TV watching patterns.

As the counterfactuals are unobserved, there is no ground truth for causal analysis in general. To validate our causalfi ndings, we need to validate the quality of matching on this user. In Fig. 6, we visualize the covariates of treated events from this user along with the covariates of matched events. In each plot, each column represents a covariate (Section 3.2), which is the profile of the user, the time of the event, and the location of the event. Each row represents an event. A quick visual inspection reveals that the plots look similar.Th e correlation coefficient between the entries of these two plots is 0.998, which indicates that they are almost identical.This shows that our matching can balance systematic biases between treated and control groups well, even on an individual user level.

To determine which level of granularity is most appropriate for user modeling, we have to summarize the effects on all users, similarly to Fig. 4(a). However, we cannot simply sum up the effects on individual users, as they may cancel each other. Consider the

following example.Th e whole population of users consists of two users with the same number of events. For user 1, the treatment increases the probability of watching *Drama* by 10%. For user 2, the treatment decreases the probability of watching *Drama* by 10%. On average, the effect of the treatment on watching *Drama* is zero. However, when the effect is viewed in the context of individual users, the treatment appears to be effective. Below we propose a metric that reflects this intuition.

Let $G = \{G_k\}_{k=1}^{K}$ be a partitioning of all events into K groups G_k, such as that each user is assigned to a single group. Let $\widetilde{\text{ATT}}_{j,y}^{k}$ be the *high-probability effect of treatment j on genre y in group k*. Then we define the gain of conditioning on groups in G as

$$\widetilde{\text{ATT}}_{j,y}^{G} = \sum_{k=1}^{K} \frac{|G_k \cap \{i : T_i = 1\}|}{\sum_{\ell=1}^{K} |G_\ell \cap \{i : T_i = 1\}|} \left|\widetilde{\text{ATT}}_{j,y}^{k}\right|, \tag{3}$$

and refer to it as the *effect of treatment j on genre y in context G*.This is a convex combination of the absolute high-probability effects in each group, weighted proportionally to the number of treated events in each group.Th e larger the number of treated events in the group, the higher the confidence on the corresponding ATT, and the higher the contribution of this group to $\widetilde{\text{ATT}}_{j,y}^{G}$.

We also define *total effect in context G* as

$$\widetilde{\text{ATT}}^{G} = \sum_{j} \sum_{y} \widetilde{\text{ATT}}_{j,y}^{G}, \tag{4}$$

which summarizes the significance of all effects, for every pair of the treatment and outcomes, given G.Th is is a strong indicator of how good the granularity of context G is for causal effects.

In Fig. 4(a), we report $\widetilde{\text{ATT}}_{j,y}^{G}$ when G is a single group, all the users in the dataset. Note that these are simply absolute effects from Fig. 3(a). In Fig. 4(b), we report $\widetilde{\text{ATT}}_{j,y}^{G}$ when each group in G is an individual user. For many genres in Fig. 4(a), the effects are significant, for instance for *Drama*, *Panel* and *News*. Differences between these two plots also reveal some trends. We observe that if we further consider smaller subgroups, such as individual users, then there are nearly no causal effects on average.Th e reason is that we do not have sufficient data for most users if considered individually. In our dataset, most users have less than 200 watching events, which means they have even less treated events. When the number of watching events is limited, we observe large ATTs but they also come with large standard errors.Th erefore, most high-probability effects in Eq. (2) are zero, and so is their convex combination in Eq. (3).

We conclude by making the following remarks. On one hand, we can observe significant effects when the groups are sufficiently large, such as the data of the whole of Australia. Although we have higher confidence in our estimated effects on these large groups, they only show overall trends for a large population, which can hardly be valuable for user-level personalization. On the other hand, if we have enough treated events for an individual user, we can observe even stronger effects. However, most of our users do not have that many watching events.Th us, there is a tradeoff between the size of the subgroups and the confidence of our estimators. In the following subsection, we show that there are other levels of granularity of context that allow us to estimate significant effects.

5.4 Causal Analysis on a Group of Users

In this subsection, we choose the most popular TV programs and conduct causal analysis of groups of people with different watching preferences. In our dataset, *Doctor Who* is the most popular TV program in *Drama* and *Peppa Pig* is the most popular in *Pre-school*. We use these two programs as the grouping criteria and build up two groups by selecting users who watched *Doctor Who* and *Peppa Pig* at least once, respectively.Th ere are 85, 252 users in the *Doctor Who* group and 25, 731 users in the *Peppa Pig* group. Only 3, 606 users watched both programs, so that the sets of users in these two groups are substantially different.

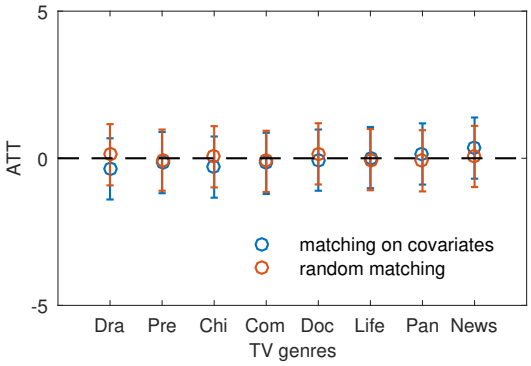
We repeat the causal analysis of Section 5.2 and Section 5.3 in these two groups.Th e obtained results in terms of the high-probability effect $\widetilde{\text{ATT}}_{j,y}$ in Eq. (2) for all treatments j and genres y are summarized in Fig. 3(b) and Fig. 3(c). Note that for each weather attribute, we have around 0.7M and 0.5M treated events in these two groups, respectively. Fig. 3 reveals some insightful trends. First, the three groups of users (i.e., the whole population in Australia, users who watched *Doctor Who*, and users who watched *Peppa Pig*) show different TV watching patterns. For example, we observe significant decrease for *Drama* when temperature is high in Fig. 3(a) but do not observe them in Fig. 3(b) and Fig. 3(c). However, there are also some consistent results observed across these groups. For example, we observe significant decrease in *Drama* when pressure is low and precipitation is high. As another example, there is no significant effect on *Pre-school* programs when weather attributes change.Th esefi ndings confirm our claims in Section 5.2 and Section 5.3 that weather attributes may be indicative of patterns in user TV watching behavior, which can be revealed through causal analyses. Second, we observe more significant effects when grouping users rather than treating them individually. Comparing Fig. 3(b) to Fig. 4(b) validates our claim that a sufficient number of treated events is needed to estimate significant effects.

Fig. 3 and Fig. 4 also provide insights on how to model user behavior. As discussed in Section 5.3, it is a challenge to balance the group size and the confidence of the estimators. In our work, we observe that we do not obtain significant effects for most of the individual users, as the watching records of a single user are limited. At the same time, the results obtained from the whole population cannot inform user-level personalization. In this subsection, we demonstrate that one way to benefit from the two worlds is to group users with similar watching preferences in order to model their TV watching patterns.

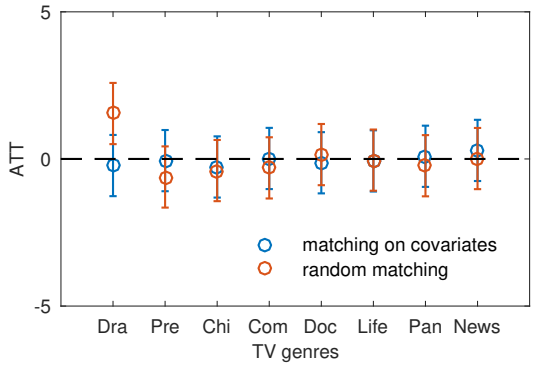
5.5 Why Causal Analysis?

In earlier sections, we showed that causal inference is a useful tool for user modeling. In this subsection, we show the necessity for causal analysis. Causal analysis and matching on covariates can correct systematic biases in data. We illustrate this by comparing two matching methods.Th efi rst method matches treated events based on covariates (see Section 3).Th e second method matches treated events randomly.

Let us consider the following example. We randomly choose 20% of events from the city of Brisbane as treated events and use the ATT in Eq. (1) to estimate causal effects. Note that the treatment is random, and therefore the true effects are zero. First, we choose



(a) Both treated and control events are from Brisbane.Th e confidence radii are the same as in Eq. (2).Th e effects are multiplied by 100 and can be interpreted as changes in the percentage of watching TV genres.



(b) Treated events are from Brisbane and control events are from the whole Australia.Th e confidence radii are the same as in Eq. (2).Th e effects are multiplied by 100 and can be interpreted as changes in the percentage of watching TV genres.

Figure 7: Comparisons of two matching methods.

control events from Brisbane and report the ATT in Fig. 7(a). We observe that both matching methods predict near-zero causal effects, which is correct.Th e reason is that the distributions of covariates, conditioned on the treatment and control, are the same.Th is shows that if the treated and control events are balanced, both matching methods work well.

Now we choose control events from the whole Australia and repeat the above experiment. Fig. 7(b) shows our results. We observe that random matching predicts higher absolute values of the ATT than matching on covariates. For instance, it estimates more than 1% increase in *Drama* and this is statistically significant. We know that this is incorrect because the true effects are zero.Th e reason is that random matching matches treatment events from Brisbane to control events from the whole Australia. Since the covariates in Australia are distributed differently from those in Brisbane, we get biases.Th us, the statistically significant increase in watching *Drama* is due to the fact that people in Australia watch less *Drama* on average (19.51%) than in Brisbane (20.89%).Th e distributions of covariates, conditioned on the treatment and control, are typically different in practice, because treatments are not assigned randomly, but rather conditionally randomly. We claim that our proposed method can correctly balance this imbalance, and therefore it should be preferred in practice for user modeling tasks.

6 CONCLUSIONS

In this paper, we study whether and how weather affects users' TV watching behavior. We conducted causal analyses using nation-scale Australian dataset and discovered interpretable causal relationships between weather conditions and users' watching behavior. We repeated the causal analysis at the granularity of individual users, but in most cases individual data turned out to be insufficient for obtaining reliable results.Th is has driven further sensitivity analyses of our approach that discovered substantial differences across subgroups of users. To the best of our knowledge, this is the first causal analysis of a large-scale dataset looking at the interplay between weather and TV watching.

Our work also raises several questions. First, in our work, our treatment variable was binary; we set hard thresholds for the "high" and "low" groups of each attribute. However, there exist works dealing with continuous treatment [12] and we posit that our dataset is rich enough for weather attributes to uncover more accurate watching patterns if these methods were used. Second, we only focused on the causal analysis and its sensitivity and did not evaluate how the learned weather dependencies can be exploited. We strongly believe these weather attributes can be incorporated into more sophisticated weather-aware recommender systems [1], and improve the quality of the generated recommendations.

It is also worth noting that the work was done in a domain-agnostic way and did not involve any meteorologists. Certain weather attributes, such as temperature and feels-like temperature or humidity and precipitation, are clearly correlated. We believe that some domain knowledge of these correlations could help us to identify the set of most influential attributes and dramatically improve our results. We also leave this work for the future.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender Systems Handbook*, pages 191–226. 2015.
[2] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci. Context relevance assessment and exploitation in mobile recommender systems. *Personal Ubiquitous Comput.*, 16(5):507–526, June 2012.
[3] A. Barker, K. Hawton, J. Fagg, and C. Jennison. Seasonal and weather factors in parasuicide. *The British Journal of Psychiatry*, 165(3):375–380, 1994.
[4] G. A. Barnett, H.-J. Chang, E. L. Fink, and W. D. Richards. Seasonality in television viewing a mathematical model of cultural processes. *Communication Research*, 18(6):755–772, 1991.
[5] M. Braunhofer and F. Ricci. *Contextual Information Elicitation in Travel Recommender Systems*, pages 579–592. Springer International Publishing, Cham, 2016.
[6] M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.
[7] E. G. Cohn. Weather and crime. *British Journal of Criminology*, 30(1):51–64, 1990.
[8] T. S. David Hirshleifer. Good day sunshine: Stock returns and the weather. *The Journal of Finance*, 58(3):1009–1032, 2003.
[9] S. Dernbach, N. Taft, J. Kurose, U. Weinsberg, C. Diot, and A. Ashkan. Cache content-selection policies for streaming video services. In *35th Annual IEEE International Conference on Computer Communications, INFOCOM 2016, San Francisco, CA, USA, April 10-14, 2016*, pages 1–9, 2016.
[10] A. Farahat and M. C. Bailey. How effective is targeted advertising? In *Proceedings of the 21st International Conference on World Wide Web*, pages 111–120. ACM, 2012.
[11] R. M. Gray and D. L. Neuhoff.Quantization. *IEEE Trans. Inform.Th eory*, 44(6):2325–29, 1998.
[12] K. Hirano and G. W. Imbens.Th e propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives.*

Wiley, 2004.
[13] G. King and R. Nielsen. Why propensity scores should not be used for matching. Working paper, 2016.
[14] Y. Koren and R. M. Bell. Advances in collaborativefi ltering. In *Recommender Systems Handbook*, pages 77–118. 2015.
[15] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.
[16] S. Li, N. Vlassis, J. Kawale, and Y. Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.
[17] D. Liang, L. Charlin, J. McInerney, and D. M. Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 951–961, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
[18] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2014.
[19] K. B. Murray, F. Di Muro, A. Finn, and P. P. Leszczyc.Th e effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, 17(6):512–520, 2010.
[20] A. G. Parsons.Th e association between daily weather and daily shopping patterns. *Australasian Marketing Journal (AMJ)*, 9(2):78–84, 2001.
[21] T. Partonen and J. Lönnqvist. Seasonal affective disorder. *The Lancet*, 352(9137):1369–1374, 1998.
[22] J. Pearl. *Causality*. Cambridge University Press, 2009.
[23] K. Roe and H. Vandebosch. Weather to view or not:Th at is the question. *European Journal of Communication*, 11(2):201–216, 1996.
[24] D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.
[25] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
[26] E. M. Saunders. Stock prices and wall street weather. *The American Economic Review*, 83(5):1337–1345, 1993.
[27] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.
[28] M. Xu, S. Berkovsky, S. Ardon, S. Triukose, A. Mahanti, and I. Koprinska. Catch-up TV recommendations: show old favourites andfi nd new ones. In *ACM Conference on Recommender Systems*, pages 285–294, 2013.
[29] S. Zong, B. Kveton, S. Berkovsky, A. Ashkan, N. Vlassis, and Z. Wen. Does weather matter?: Causal analysis of TV logs. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 883–884, 2017.