Causal Inference for ATM Counterfactual Estimation

Akhil Shah
RAND Corporation
Santa Monica, CA 90407
Email: ashah@rand.org

Abstract—**We propose that the Rubin potential outcomes framework of causal inference can be used to statistically estimate counterfactuals, by definition never observable, of Traffic Flow Management Initiatives (TFMI), as a novel means of quantifying the performance of Air Traffic Management (ATM) actions, despite confounding factors. Specifically, we apply the method of Propensity Scores to estimate counterfactuals and compute the increase in hourly average airborne delay which would have resulted without a Ground Delay Program, using an eleven month span of hourly weather, traffic, and delay data at JFK. Our introduction also summarizes the concepts of causal inference required for our analysis. We also offer suggestions to improve and extend our initial application of casual inference. Technical details of propensity score modeling are further covered in an appendix.**

Keywords—**ATM Performance Measurement, Causal Inference, Propensity Scores**

I. Introduction

Traffic managers at the Air Traffic Control System Command Center (ATCSCC) and other stakeholders, such the Airlines Operations Centers (AOC), should be able to estimate performance metrics, such as ground and airborne delays, due to various courses of actions. Moreover, improvements in Air Traffic Management (ATM) require accurate estimates of actual and potential outcomes of specific Traffic Flow Management Initiatives (TFMI), interventions which ensure and enhance performance of the National Air Space, in order to enable comparison of alternative courses of actions. These performance estimates of alternative courses of action can result from simulation or statistical inference. Thus motivated, we present a novel application of a well developed statistical methodology, causal inference, which enables estimating the effectiveness of interventions, such as TFMI, through constructing counterfactual outcomes, by definition never observable, and has been useful in various other domains where random trials are impractical, including health [1], [2], economics [3], [4] and ecological contexts[1]. However, to the best of our knowledge, casual inference has not yet been explored in the ATM research community. We argue that causal inference may provide a statistically rigorous framework to estimate the effectiveness of TFMI, despite the presence of *confounding* in ATM datasets.

In addition to summarizing the conceptual aspects of causal inference, we will illustrate a specific ATM use case: estimating the potential outcome of not implementing a particular TFMI, namely the potential airborne delay which if a Ground Delay Programs (GDP) had not been implemented during hours when it was in fact implemented at JFK.

A. Confounding in ATM Data

Consider the case of evaluating the efficacy of a medical treatment. One option is to implement randomized control trials, where units are randomly assigned to treatment and control groups, and is sometimes feasible. However randomized control trials, considered the gold-standard to statistically evaluate the effectiveness of a treatment [5], are not always possible to implement. For example, lets consider the efficacy of smoking cessation counseling for smokers admitted to a hospital for a heart attack [1]. In particular, we are interested in the following question: does smoking cessation counseling, prior to discharge from the hospital, increase the lifespan of smokers who have suffered a heart attack? If a randomized control trial were possible in this situation, then the usual methods of regression would suffice to answer this question statistically. However there are various barriers to voluntary participation and completion of treatment [1], and thus a random controlled experiment is not possible in this example. Statisticians call such a situation an “observational study,” and commonly, there are systematic differences between patients who receive treatment and those who do not, which must be accounted for in a sound methodological manner when assessing the effect of a treatment on the outcome of interest (e.g. mortality).

Similarly, in the ATM context, it is infeasible to conduct randomized control trials. For a particular TFMI, such as a Ground Delay Program (GDP), blocks of time or individual flights are clearly not assigned randomly to the initiative, but are the consequence of a collaborative decision making process based on various factors such as weather, traffic, and capacity. Furthermore, outcomes, such as airborne delay are also influenced by these same covariates; hours of inclement weather are more likely to be assigned GDP and they are also likely to experience higher average airborne delay than hours with clear weather. Thus to empirically estimate the effectiveness of a TFMI like GDP requires accounting for these confounding factors.

Estimates of performance metrics can enable tradeoffs between different courses of action. More specifically, counter-

[1] For example: https://www3.epa.gov/caddis/da_advanced_5.html

factual estimates of performance due to a given TFMI, can provide decision makers an estimate of the potential cost of inaction, for example average airborne delays if a GDP had not been imposed at a given arrival airport. These counterfactual estimates may be useful both at operational and strategic levels when evaluating the cost and benefits of various TFMI at various spatial and temporal scales. However estimating counterfactual outcomes is methodologically challenging in the ATM context where the notion of random trials or random assignment of flights to is impractical. We propose to employ the statistical methodology of causal inference to derive these counterfactual estimates, which have the advantage of being derived from a statistically rigorous framework that has been applied in various other domains, such as health, education, and economics. To our knowledge, this is the first application of causal inference to ATM.

As explained in [5], the fundamental difficulty of observational studies versus random controlled trials, is the presence of *confounding*: the outcome of interest (e.g. mortality) is influenced by both the treatment status (i.e. whether the patient received or did not receive treatment) and the baseline characteristics, which are often systematically different between the treatment and control groups. Although there exist regression adjustment techniques that attempt to account for confounding, there are many reasons for which they are not robust [5], [1], [6]. An alternative method to eliminate or reduce confounding uses the idea of propensity scores and the potential outcomes framework on which they are based.

B. *Related research*

Various approaches can be used to analyze alternative courses of actions for ATM. One possibility is to use a dynamic systems modeling approach, employing a combination of analytic models and discrete-event simulations. Tools such as ACES and FACET [7] can be used to simulate airborne delay statistics from various TFMI implementation strategies with realistic traffic flows. Queuing theoretic models, with varying degrees of simplifications, can also be used to to analyze delay statistics resulting from a range of GDP implementation parameters [8], [9], [10], [11], [12].

Alternatively one can use techniques that don't directly model the dynamics of weather and traffic in the NAS through analytic or discrete-event approaches, but rather through statistical techniques for producing *counterfactual* scenarios unobserved in the historical data. These counterfactual scenarios can then be used to estimate the impact of potential TFMI given weather and traffic forecasts and thus aid 'what-if' analysis required of decision makers. For example [13] uses a statistical simulation technique ("quantile equivalence") to generate counterfactual scenarios of demand and throughput at LGA, EWR, and JFK, and consequently predicts delay at these airports. This method does account for a single, but relevant, weather feature using an empirical non-parametric procedure to statistically simulate counterfactual scenarios by mixing *observed* throughput and demand in various time periods for a given weather condition (either VMC or IMC).

C. *Outline*

In this report we will examine another approach, Causal Inference, to generate counterfactual estimates, which can additionally quantify and reduce confounding, in order to attribute effectiveness of interventions (like TFMI) in terms of relevant outcomes. Causal Inference actually comprises several statistical methods, including propensity scores [1], which we will specifically employ for counterfactual estimation to evaluate TFMI impact on system outcomes in the presence of confounding factors such as weather and demand.

Our initial examination of airborne delay under a GDP is the simplest application of Causal Inference, using hourly data as the unit of analysis, with binary treatment assignment (GDP or no GDP). In our conclusions, we offer the ATM research community various possible extensions of our initial application of Causal Inference. We will briefly discuss more sophisticated possibilities employing other units of analysis, such as individual flights, multiple discrete treatments for GDP across multiple airports, and continuous treatment variables such as the amount of ground delay. We will also articulate the various challenges which accompany such an undertaking in our conclusion.

II. Essential Concepts of Causal Inference

A. *Potential Outcomes*

The Rubin potential outcomes framework [14] imagines two possible treatment assignments for each unit of analysis (e.g. a patient, a flight, or a block of time), i.e. a treatment and control, and denotes the treatment status for each unit with the indicator variable Z ($Z = 0$ for control and $Z = 1$ for treatment). For each unit, the effect of the treatment on the outcome (e.g. mortality, delay, etc.) Y is defined to be $Y_i(Z = 1) - Y_i(Z = 0)$. Notice however that for each unit, only one reality is observed, and thus to compute effect, we must be able to statistically estimate the counterfactual or potential outcome. For example, for a subject who ultimately receives treatment, we can only observe $Y_i(Z = 1)$, but not the counterfactual $Y_i(Z = 0)$. The potential outcomes framework attempts to estimate counterfactuals so that the average effect of the treatment can be computed either for all subjects, called the average treatment effect (ATE), or for only the subject that received treatment, called the average treatment effect on the treated (ATT). Note that the analyst must decide which quantity is more appropriate to estimate; for example in the smoking cessation counseling example, ATT is the appropriate quantity to estimate as it is not realistic that all patients would likely elect treatment [1]. However if the treatment were instead a brochure on smoking, the barrier to treatment entry is low, and thus ATE would be appropriate to estimate as it is realistic to assume that all subjects could potentially be part of either treatment or control group. We argue that in our context of TFMI "treatments" (e.g. GDP), ATT is the more appropriate quantity to estimate, because it is unrealistic to assume that all time-periods (e.g. even those with "good" weather and normal traffic and capacity characteristics) would potentially be subject to TFMI action. Also note that to calculate ATT, $E[Y(Z=1) - Y(Z=0)|Z=1]$, we only need to estimate the counterfactual for units in the treatment group (i.e. estimate the counterfactual Y(Z=0) for the treated subjects), whereas for ATE, the counterfactual for the control group must also be estimated.

In the next section we explain how confounding can be reduced by balancing the baseline characteristics using the propensity score. The aim of balancing pretreatment covariates can also be viewed as transforming data from an observational study so it resembles the gold standard of a randomized control trial [5].

B. Propensity Score

The propensity score is defined as $e_i = Pr(Z_i = 1|X_i)$, namely the *probability* that subject i with baseline characteristics described by the covariate vector X_i is assigned to the treatment group. Note that all subjects have a propensity score in the potential outcomes framework, regardless of whether they were actually in the treatment or control group. The important statistical property of the propensity score is that it is a *balancing score* [5]: conditional on the propensity score, the distribution of baseline covariates is similar between treated and control subjects. Thus for a set of subjects with the same propensity score (value of e_i), there should be no statistically significant difference in baseline covariates, and thus a counterfactual outcome can be estimated, allowing the eventual estimate of ATT or ATE.

Thus far we have summarized the fundamental obstruction to causal analysis in observational studies, namely confounding, and have also reviewed how the potential outcomes framework and counterfactual estimation can be used in principal to overcome confounding, and how the propensity score's balancing property can provide such counterfactual estimation[2]. In the appendix, we clarify the mechanics of how propensity scores are estimated, namely the various model and the model fitting procedures, and how the scores are then used to balance covariates and estimate ATE or ATT using various methods.

Our application of the potential outcomes framework using propensity scores to estimate the impact of potential TFMI uses the following analogy: each record is an hour time period at a given airport; measured baseline characteristics are historical forecasts of weather (relevant features from TAF) and traffic (hourly arrival data from ASPM); treatment assignment is the occurrence of a TFMI in the time period (such as GDP); and measured outcome is the hourly averaged airborne delay, also recorded in ASPM.

III. Application to GDP

TFMI are designed to increase the safety and efficiency of the nation's air transportation system, and are necessary during inclement weather and in other situations where demand exceeds capacity. When the arrival capacity of airport cannot accommodate demand, either due to weather induced diminished capacity or volume induced enhanced demand, a GDP is often implemented. A GDP will purposefully delay flights on the ground at the various origin airports to avoid more costly airborne delays that would be incurred at the arrival airport[3]. Designing a GDP involves setting a planned airport acceptance rate (PAAR), commensurate with the forecasted arrival capacity, and then sequencing affected flights with controlled departure times, ideally with an equitable distribution of delays [8]. Increasing or decreasing this PAAR, or inversely the inter-arrival time between flights, transfers the overall delay of all flights subject to the GDP between airborne or ground portions (higher PAAR means more of the overall delay is absorbed in the air).

Our causal analysis will specifically focus on estimating the counterfactual outcomes of a GDP at JFK, using a variety of data sources, to be detailed shortly, that span eleven months between October 2013 and September 2014. As noticed in [16], the average airborne delay for flights with GDP ground-delays is roughly four minutes larger than the average airborne delays for flights that did not receive an EDCT under a GDP. However it is commonly understood that in absence of a GDP, namely the counterfactual of not applying ground-delay to the affected flights, the arrival airport would experience even larger airborne delays than actually observed in reality. In terms of a queuing model [8], not instituting a GDP during diminished arrival capacity or enhanced traffic volume, would correspond to an aggressive policy with a PAAR exceeding the actual arrival capacity. Although such a policy may increase airport utilization it would incur the cost of larger airborne delays.

In one sense, our present analysis attempts to quantify this potential, yet unobserved, savings in airborne delay from implementing a GDP. As the counterfactual is not actually observed for the ground-delayed flights, we require a statistically sound methodology to estimate it from the historical recorded data, which takes into account the systematically different weather and traffic conditions faced by flights during the presence and absence of a GDP. Although analysis of historical records of TFMI can in principal demonstrate the relative merits of courses of action and their affect on the observed outcomes, there is a fundamental challenge of accurately accounting for the distinct conditions, such as weather and demand, experienced during previous time periods, and the degree to which differences in outcomes such as delays, are also attributable to these *confounding* conditions.

IV. Data Analysis

Consider estimating the effectiveness of GDP applied at JFK in terms of one of its intended outcomes, reduction in airborne delay. We choose the unit of analysis to be an hour block of time at the arrival airport and examined 11 months of hourly

[2] See [5] for the further discussion on statistical assumptions that underly the balancing property of propensity scores and which allow confounding to be reduced or eliminated

[3] Further examples of TFMI and detailed descriptions can be found in [15].

data between October 2013 and December 2014. For each analysis unit we extract whether there was a GDP en-force at that hour as the (binary) treatment variable, relevant weather and traffic covariates, and the observed outcome of interest, namely the airborne delay.

If instead of GDP, we were analyzing other TFMI with causal inference, other covariate (e.g. weather), treatment (when and how TFMI are implemented) and outcome (intended impacts) datasets may be required. For example, in the case of Reroutes, weather features derived from measurements of convective cloud top altitude were found to be relevant covariates [19]. A fuller discussion of causal inference applied to other TFMI is left for future research.

A. Extracting Treatment Status

To determine the exact hours when the treatment was applied, or equivalently, which hours GDP were in effect at JFK, we used the National Traffic Management Log (NTML), which records when GDP are announced, started, revised, and finally cancelled. To discern which hours within the analysis period contained a GDP, we parsed the NTML data by arrival airport and tracked the start and stop times of GDP that were initiated. Furthermore, as is common, some of these initiated GDP were modified, and thus required that we track each "root advisory" (initial GDP) through its event history and note actual start and stop times rather than those announced initially. We note that of the 117 GDP root advisories in our data, 18 of those were implemented with a start time that preceded the send time (time that the advisory was transmitted as recorded in the NTML dataset); namely these GDP were 'backdated', and represent a reactive decision with a lag of 20 minutes or less. In order to separate the effects of weather forecast uncertainty from our causal analysis, we will focus on these reactive decisions where weather is more likely certain, and thus only use the hours which had backdated GDP (treatment group) or no GDP (control group) in our causal analysis. Our total sample size for the chosen unit of analysis of hour blocks in the eleven month period are thus: $N_T = 84$, and the number of hours in the control group $N_C = 6519$.

B. Extracting Covariates

The propensity to apply the treatment to the unit, or equivalently, the likelihood of implementing a GDP at a given hour for arrivals at JFK, will depend on various covariates such as weather and traffic volume at the arrival airport. Both forecasted and observed weather variables will determine such a propensity to implement a GDP. In addition the scheduled number of arrivals, the presently observed arrival queue will be relevant to model the propensity score for GDP in a given hour. We use Terminal Aerodrome Forecasts (TAF) for forecasts of weather at the arrival airport, and also use the FAA's Aviation Systems Performance Metrics (ASPM) database for observed hourly weather. In addition we use ASPM for scheduled arrivals, quarter-hour arrival demand and actual landings.

C. Extracting Observed Outcome

Finally, for the outcome of interest, namely airborne delay, we again used ASPM data which records hourly average airborne delay. Specifically, the ASPM Airport Analysis module (publicly available data) records hourly airborne delay averaged over flights landing at a selected airport at that hour. If the unit of analysis were individual flights, one would require individual flight data from ASPM. In future analysis we will apply causal inference using individual flights as the unit of analysis, and as such, we have noted that the reported airborne delay is based on scheduled or flight-plan estimated time en-route (ETE). As noted by other researchers [17], [18], schedule padding is a common airline practice which can exaggerate actual ETE. A more robust method to estimate "nominal" airborne delay would be to group flights according to the triplet of origin, destination, and carrier to extract the tenth percentile of actual (observed) en-route times. However such a computation requires access to individual flight data, which was not available to us during our analysis. Using the percentile method[4] to extract "nominal" ETE from individual flight data to directly compute hourly averaged airborne delay would be interesting for comparison to the results presented herein.

D. Covariate Imbalance for GDP Treatment Propensity

Previous research has highlighted meteorological conditions such as visibility, windspeed, crosswind, and cloud ceiling as relevant predictors of GDP [20], especially for the New York metro region and including JFK. ASPM data provides hourly observed values for each of these weather covariates at the arrival airport, whereas we use the most current (shortest horizon) TAF for forecasted values of these variables (except for cloud ceiling). Table I shows the mean values for these forecasted and observed (denoted with "_ASPM") meteorological covariates averaged overall all hours for which a GDP was present, denoted as E(Y1|t=1), and for all hours in which a GDP was absent, denoted E(Y0|t=0). Recall this notation arises from the Rubin potential outcome framework explained earlier. Also note, that for each covariate, the difference between the GDP and non-GDP hour, or equivalently the treatment and control groups, are statistically significant as determined by the low p-values[5]. This is statistical evidence of covariate imbalance, which must thus be adjusted prior to estimating the counterfactual airborne delay and ATT.

Also shown in Table I are the scheduled number of hourly arrivals at JFK ("A_JFK") and approximate length of arrival queue ("qlength"): the difference between number of flights demanding arrival (ARRDEMAND in ASPM) and number of flights which actually arrived or landed (EFFARR). ARRDEMAND[6] and EFFARR are calculated on a quarter hour

[4] See [18] for details and justification of this method.

[5] Usually, p-values less than a given significance threshold are considered evidence to reject the null-hypothesis, namely reject the hypothesis that the samples were drawn from the same population.

[6] See ASPM for further details on how demand units are computed http://aspmhelp.faa.gov/index.php/SAER

basis and we have aggregated them to the hour for our unit of analysis.

	E(Y1|t=1)	E(Y0|t=0)	p-value
qlength	53.6	2.3	0.00
A_JFK	34.8	21.9	0.00
windspeed	12.3	10.9	0.08
visibility	7.2	9.3	0.00
crosswinds	9.7	7.5	0.00
vis_ASPM	7.7	9.3	0.00
windspeed_ASPM	15.2	11.4	0.00
ceil_ASPM	228.0	481.7	0.00
crosswind_ASPM	8.6	7.0	0.00

TABLE I

COVARIATE IMBALANCE BETWEEN AVERAGES OF TREATMENT (GDP HOURS, $E(Y1|t = 1)$) AND CONTROL (NON-GDP HOURS, $E(Y0|t = 0)$) GROUPS FOR VARIOUS WEATHER AND TRAFFIC COVARIATES (PREDICTORS) RELEVANT FOR GDP DECISIONS. ALMOST ALL COVARIATE DIFFERENCES ARE STATISTICALLY SIGNIFICANT AS DETERMINED BY THE P-VALUE.

E. Propensity Score Estimation and Balance Improvement

We then estimated the propensity scores $e_i \equiv Pr(Z_i = 1|\mathbf{X}_i)$ for each time period i, using a Generalized Boosted Model (GBM) [21] based on the weather and traffic covariates for each time period $\mathbf{X}_i$ described above. Note that the process of generating a propensity score is very similar to supervised learning[7] As in any model fitting procedure, one requires a criterion to pick the best model. In casual inference, the best propensity score model is one that achieves the best balance between covariates and not necessarily the model which most accurately predicts treatment. The propensity scores from the optimal covariate balance model can then be used to generate weights $w_i \equiv 1/e_i$ to estimate the counterfactual airborne delay for a given forecast of weather and traffic, a procedure generally called Inverse Probability of Treatment Weighting (IPTW) [5]. See the appendix for further details about propensity score modeling and the IPTW methodology.

Using open source software [21], we estimated the best propensity score model and after applying the IPTW weights to the covariates, we notice an improvement in covariate balance as shown in table II. Notice there are no longer any statistically significant difference between the treatment E(Y1|t=1) and weighted control group E(Y0|t=1), again using the notation of the potential outcomes framework. As an additional check on the propensity score model, we graphically assess the overlap of estimated propensity scores between the treatment and control groups from the resulting model. Note that zero overlap of propensity scores would mean that one cannot use the control groups outcomes to generate counterfactual outcomes for the treatment group (see appendix for further discussion) [1].

	E(Y1|t=1)	**E(Y0|t=1)**	p-value
qlength	53.6	44.5	0.09
A_JFK	34.8	33.2	0.61
windspeed	12.3	12.5	0.76
visibility	7.2	7.9	0.11
crosswinds	9.7	9.3	0.73
vis_ASPM	7.7	7.7	0.87
windspeed_ASPM	15.2	14.3	0.72
ceil_ASPM	228.0	265.8	0.16
crosswind_ASPM	8.6	8.6	0.73

TABLE II

IMPROVED COVARIATE BALANCE USING IPTW TO ESTIMATE THE COUNTERFACTUAL CONDITIONS $E(Y0|t = 1)$. NOTE THAT THE LARGE P-VALUES INDICATE DIFFERENCES IN MEANS ARE NOT STATISTICALLY SIGNIFICANT.

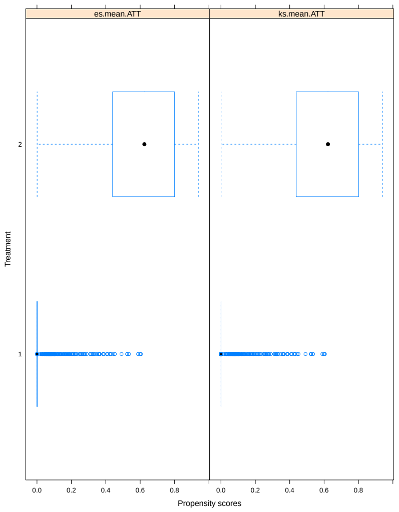

Fig. 1. Overlap of the estimated propensity scores between the non-GDP hours or control group (1 on y-axis) and the GDP hours treatment group (2 on the y-axis). There is non-zero overlap as required for counterfactual estimation, and in particular there are hours in the control group that have propensity scores similar to the treatment group, implying similar weather and traffic covariates, and thus will be weighted more greatly by IPTW for estimating the counterfactual. The subplots represent two different propensity score models fit with different criteria for optimal covariate balance (see appendix).

F. Outcome Analysis

As we have determined that our propensity score model leads to greater covariate balance as show in table II, we can continue onwards to analyze the effect of treatment (GDP) on the outcome (airborne holding delay). Since the two groups are now balanced, the situation is similar to a RCT, and one can linearly regress the outcome simply on an indicator variable for treatment. Note that computationally one uses weighted linear regression with IPTW weights computed from the previously fitted propensity score models [21]. After using this procedure we estimate the ATT, E(Y1|t=1)-E(Y1|t=0), to be -1.4 minutes. In other words, had a GDP not been applied to the hours at JFK when it in fact was, the hourly average airborne delay would increase by 1.4 minutes (with a standard error of the same magnitude). Although consistent with the common understanding that GDP reduces airborne

[7] More precisely, if the supervised learning method employed a soft-decision threshold, then the propensity score would be produced as an intermediate step by the classifier when estimating whether a given record should be assigned a label of GDP or "no-GDP" based on its weather and traffic feature vector $\mathbf{X}_i$.

delay, our results also indicate this difference is not statistically significant. To further analyze this results, one can enlarge the sample size (recall we focused only on backdated GDP to avoid weather forecast uncertainty), but should also consider the possibility that GDP affected flights may receive additional en route airborne delay from additional TFMI [16].

V. Conclusion

Our simplified application of causal inference served to illustrate a novel methodology to the ATM research community. However in order to further validate results from such an analysis, namely counterfactual estimation of airborne delays from GDP, it would be useful to compare against simulations or queuing models.

As demonstrated with a simplified constant capacity (service rate) queuing model in [8], the mean Airborne delay is a function of the ratio of PAAR and the arrival capacity, subject to random fluctuations due to demand uncertainties, which increases when the PAAR exceeds the actual AAR[8]. A direct comparison to such a queuing model would require casual analysis not with a binary treatment variable as we have, but rather a continuous treatment variable using empirical data to form the ratio of PAAR and arrival capacity (both available in ASPM). We leave such modeling for future research, which can leverage more advanced propensity score estimation techniques for continuous treatment[9], which are relatively new. A further anticipated complication with comparison to queuing models is that arrival capacity is also both dynamic and stochastic, while also influenced by departure demand [23].

Furthermore, although we have chosen our unit of analysis to be an hour block of time, other units of analysis may also be possible. For example, if one had access to flight level data, in terms of both outcomes (airborne delay) and treatment assignment (GDP induced ground delay), it may be feasible to consider a single flight as the unit of analysis. However again, the challenge with individual flight data will be to account for possibly additional TFMI that may increase the airborne delay to flights already subject to GDP, as hypothesized in [16]. It may be difficult to account for en-route airborne delays, some of which may not be due TFMI at all, but rather due to Air Traffic Control actions, and thus requiring detailed radar data for analysis of actual flight trajectories [24]. Determining if there is a systematic difference for GDP and non-GDP affect flights from these sources of additional airborne delays is a precursor to causal analysis that can account for such an imbalance to estimate counterfactual airborne delays.

Even in our simplified setting of binary treatments, there are considerations that deserve further analysis. Causal inference requires certain assumption be satisfied, including that all confounding variables that affect treatment and assignment have been measured [5]. Our modeling took into account certain meteorological and traffic covariates (I), but it is well known that GDP can be implemented for other reasons, including maintenance events or large-scale disruptions to the NAS. To what degree such an assumption is satisfied needs further investigation. Another assumption of causal inference is that the potential outcome of a unit be unaffected by the treatment assignment of other units[10]. In our case such an assumption would require that the hourly average airborne delay be independent of GDP assignment of other hour blocks. However there can be autocorrelation of airborne delay in neighboring hour blocks, based on whether those neighboring blocks are GDP or non-GDP hours. Further analysis of such treatment dependent autocorrelation is required and may necessitate aggregation to larger blocks of hours or an entire day to satisfy the necessary assumption. Such time-varying treatment assignment may also necessitate more advance Causal Inference methodology, which have been recently developed[25].

There also several possible extensions of our initial application of causal inference framework and methodology to ATM. One is to simply consider how such a framework may apply to other TFMI besides GDP, using the language of treatment, outcomes, and propensity score analysis. Further extensions to the GDP context could include hourly analysis of the entire New York metro, using weather covariates from all airports simultaneously in multi-valued (but not continuous) treatment setting, where treatment corresponds to a categorical variable indicating GDP absence or presence at each of the major airports. Another extension is to consider continuous valued treatments by explicitly considering the length of the ground-delay as the treatment variable, and estimating what in statistics is called the "dose-response curve" [22], with airborne delay playing the role of the response.

VI. Appendix

A model for the propensity score is a function from the space of covariates X_i to $0 < e_i = Pr(Z_i = 1|X_i) < 1$, or more traditionally to the log-odds e_i, namely:

$$\log \frac{e_i}{1 - e_i} = F(X_i) \qquad (1)$$

The most basic model for $F(X)$ is to assume linearity, $F(X) = \beta X$, which is then fit just as linear logistic regression models are[11]. However this simplest linear model has been shown in simulation and actual studies to not achieve the best balance between treatment and control covariates .

We note here that previous ATM research has also used linear logistic regression to model the probability of a GDP occurring, with a goal of fitting the most accurate GDP classifier to ultimately identify similar weather impacted airport days [20]. However we emphasize that our goal is *not* to derive

[8] More complex queueing models of airborne delay are possible, as in [11], which account for details of the three-dimensional arrival flows using radar data and subsequently employ non-homogenous queueing models to allow for temporal variations in arrival and service rates. Non-Homogenous stochastic processes have been proposed as offering a better fit to observed airborne delay statistics [12] when compared to the simpler, but often used, assumption of a Poissionian arrival process.

[9] See for example [22] and references therein.

[10] This is called the stable unit treatment value assumption (SUTVA).

[11] Notice that we are regressing the covariates X_i on the log-odds of the probability of treatment e_i, and *not* on the outcomes directly.

the most accurate classifier but instead to use the probability of a GDP (or other TFMI) occurring as a balancing score for counterfactual estimation, and thus even if we employed linear logistic regression, the optimal model coefficients obtained using metrics for balance, would certainly be different that those using metrics for accurate classification.

More robust alternatives to linear logistic models for propensity score include machine learning methods [26], such as Generalized Boosted Models (GBM), which employ combinations of non-parametric piecewise-linear functions that adapt to the data and are thus more flexible than a linear model. In addition GBM is implemented in open-source statistical software [27] and can thus be easily replicated by other researchers. Furthermore, when GBM is used as the model for propensity score, fitting procedures which employ optimization to tune these piecewise linear functions to achieve best balance between treatment and control covariates are also readily implemented in open-source statistical software [21]. Furthermore there are various quantitative balance metrics that can be easily accessed to assess the quality of the resultant propensity score model [21].

Once the propensity score model has been fit, one can use the resulting propensity scores for each subject, e_i, to balance the covariates X_i between treatment and control groups and thereby reduce or eliminate confounding. The four principal methods to reduce confounding using propensity scores are: matching, stratification, inverse probability of treatment weighting (IPTW), and covariate adjustment [5]. We will only summarize IPTW as it has been thoroughly implemented and tested in software [21], and has also been extended to multiple treatments [6], which will eventually be required if we want to consider the effect of various TFMI options beyond just "GDP or no-GDP." Previous research on TFMI [28] has shown there are likely many categories of TFMI that occur and which combine the various courses of actions available to decision makers, each with their own specific operational parameters. Thus extensions of propensity score modeling beyond binary treatments is a desirable property of IPTW.

Recall that Z_i is an indicator variable which denotes treatment status, i.e. $Z_i = 1$ if the subject received treatment. IPTW defines weights for each subject that capture the inverse probability the subject received treatment as follows [5]:

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}. \tag{2}$$

The intuition behind the weights is the following: those subject in the control group ($Z_i = 0$) whose propensity scores (probability of being selected for treatment) are relatively higher (e_i is larger) are "weighted up" and thus their covariates are more greatly weighted when assessing balance after weighting. Various balance measures after weighting with w_i are possible using significance testing for differences of means, medians, variance, and Kolmogorov-Smirnov statistics[21].

These weights can also be used to adjust the outcome for each subject Y_i and thereby simulate counterfactuals used to estimate ATT or ATE. For example to estimate ATT, one weights the outcomes (e.g. airborne delay) for the treatment group with unity and for the control group with weights $w_i = e_i/(1 - e_i)$. Then the ATT treatment effect $E(Y_i(Z_i = 1) - Y_i(Z_i = 0)|Z_i = 1)$ can be estimated by regressing on a single variable, the treatment indicator [21], or in the simplest model by arithmetic mean of the weighted counterfactual outcome.

A. Advantages of propensity score over regression to reduce confounding

As explained in [6], there are several reasons why propensity score techniques are advantageous over such regression-based techniques and here we simply summarize these advantages:

- *dimensional reduction*: propensity scores summarize all covariates into a single score and act as an important dimensional reduction tool for evaluating treatment effects. Whereas regression methods require specifying a model that depends on all covariates (and various interactions).
- *grounded in rigorous framework*: propensity score methods derive from a formal statistical model for causal inference, the potential outcomes framework, so that causal questions can be well-defined and explicitly specified and not conflated with the modeling approach as they are with traditional regression approaches
- *robust against model misspecifcation*: propensity score methods do not require modeling the mean for the outcome, which can help avoid bias from misspecification of that model
- *avoid extrapolation*: propensity score methods avoid extrapolating beyond the observed data unlike parametric regression modeling for outcomes which extrapolate whenever the treatment and control groups are disparate on pretreatment variables
- *propensity score adjustments* (e.g. IPTW weights) can be determined using only the pretreatment covariates and treatment assignments, eliminating the influence that estimated treatment effect can have on model specification of covariates.

Acknowledgment

The author would like to thank Heather Arneson, Paul Lee, Anthony Evans, and Karl Bilimoria of NASA Ames for useful correspondence and feedback. We also thank our colleagues Matthew Cefalu and Beth Ann Griffin of the RAND Center for Causal Inference for fruitful discussions. In addition we thank Kenneth Kuhn, Christopher Skeels, and Kyle Murray for collaboration on related research.

References

[1] P. C. Austin, "A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality," *Multivariate behavioral research*, vol. 46, no. 1, pp. 119–151, 2011.

[2] C. G. Victora, J.-P. Habicht, and J. Bryce, "Evidence-based public health: moving beyond randomized trials," *American journal of public health*, vol. 94, no. 3, pp. 400–405, 2004.

[3] L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207–3260, 2013.

[4] M. B. Tariq, M. Motiwala, N. Feamster, and M. Ammar, "Detecting network neutrality violations with causal inference," in *Proceedings of the 5th international conference on Emerging networking experiments and technologies*. ACM, 2009, pp. 289–300.

[5] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.

[6] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette, "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in medicine*, vol. 32, no. 19, pp. 3388–3414, 2013.

[7] B. Sridhar, G. Broto Chatterji, and S. Randall Grabbe, "Benefits of direct-to tool in national airspace system," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 1, no. 4, pp. 190–198, Dec 2000.

[8] M. Ball, T. Vossen, and R. Hoffman, "Analysis of demand uncertainty effects in ground delay programs," in *4th USA/Europe Air Traffic Management R&D Seminar*, 2001, pp. 51–60.

[9] J. Kim, M. Tandale, and P. Menon, "Air-traffic uncertainty models for queuing analysis," in *9th AIAA Aviation Technology, Integration, and Operations Conference*, 2009.

[10] P. Sengupta, M. D. Tandale, and P. Menon, "Computational queuing analysis: An application to traffic flow analysis of the nas," in *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, 2010, p. 9137.

[11] C. Gwiggner and S. Nagaoka, "Data and queueing analysis of a japanese air-traffic flow," *European Journal of Operational Research*, vol. 235, no. 1, pp. 265–275, 2014.

[12] M. V. Caccavale, A. Iovanella, C. Lancia, G. Lulli, and B. Scoppola, "A model of inbound air traffic: The application to heathrow airport," *Journal of Air Transport Management*, vol. 34, pp. 116–122, 2014.

[13] A. Kim and M. Hansen, "Deconstructing delay: A non-parametric approach to analyzing delay changes in single server queuing systems," *Transportation Research Part B: Methodological*, vol. 58, pp. 119–133, 2013.

[14] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[15] FAA, "Tfm in the nas – info for flight ops personnel," *ATCSCC Products*, 2009. [Online]. Available: https://www.fly.faa.gov/Products/Training/Traffic_Management_for_Pilots/TFM_in_the_NAS_Booklet_ca10.pdf

[16] K. Bilimoria, "Analysis of additional delays experienced by flights subject to ground holding," in *AIAA Aviation Technology, Integration, and Operations Conference*, 2016.

[17] L. Hao and M. Hansen, "How airlines set scheduled block times," in *10th USA/Europe Air Traffic Management Research and Development Seminar, Chicago IL*, 2013.

[18] G. Skaltsas, "Analysis of airline schedule padding on us domestic routes," Ph.D. dissertation, Massachusetts Institute of Technology, 2011.

[19] H. Arneson, "Initial analysis of and predictive model development for weather reroute advisory use," in *15th AIAA Aviation Technology, Integration, and Operations Conference*, 2015, p. 3395.

[20] S. R. Grabbe, B. Sridhar, and A. Mukherjee, *Clustering Days with Similar Airport Weather Conditions*. American Institute of Aeronautics and Astronautics, 2016/02/19 2014. [Online]. Available: http://dx.doi.org/10.2514/6.2014-2712

[21] G. Ridgeway, D. McCaffrey, A. Morral, L. Burgette, and B. A. Griffin, "Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package," *R vignette. RAND*, 2015.

[22] N. Kreif, R. Grieve, I. Díaz, and D. Harrison, "Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury," *Health economics*, vol. 24, no. 9, pp. 1213–1228, 2015.

[23] A. Jacquillat and A. R. Odoni, "Endogenous control of service rates in stochastic and dynamic queuing models of airport congestion," *Transportation Research Part E: Logistics and Transportation Review*, vol. 73, pp. 133–151, 2015.

[24] S. Belkoura, J. M. Peña, and M. Zanin, "Generation and recovery of airborne delays in air transport," *Transportation Research Part C: Emerging Technologies*, vol. 69, pp. 436–450, 2016.

[25] K. Imai and M. Ratkovic, "Robust estimation of inverse probability weights for marginal structural models," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 1013–1023, 2015.

[26] B. K. Lee, J. Lessler, and E. A. Stuart, "Improving propensity score weighting using machine learning," *Statistics in medicine*, vol. 29, no. 3, pp. 337–346, 2010.

[27] G. Ridgeway *et al.*, "gbm: Generalized boosted regression models," *R package version*, vol. 1, no. 3, 2006.

[28] K. D. Kuhn, A. Shah, and C. Skeels, "Traffic flow management plan characterization," *NASA Research Report*, 2016.

Akhil Shah is a Senior Engineer with the RAND Corporation where he develops machine learning and statical methodology for air-traffic and cybersecurity related research problems. He received his PhD in Theoretical High Energy Physics and M.S. in Electrical Engineering from UCLA. His previous research involved communication system design and analysis for both wireless and optical applications.