In this paper, we use real-world data [1] collected over two years from operational cellular networks (GSM, UMTS and LTE) and demonstrate that external factors such as seasonality (because of foliage), weather changes (strong winds, rainfall, thunderstorms or extreme events like hurricanes or tornadoes), traffic changes (due to holidays, major events) and network events (outages and other maintenance activities) have a significant impact on service performance. Fig. 1 shows an example how strong winds in a region negatively impacted service performance (increase in the dropped voice call ratios) and this coincidentally co-occurred with the time of a configuration change at a network element. Anyone assessing the performance impact of the change (either manually or using an automated tool) without the knowledge of the current weather conditions and its performance impact, would have made incorrect conclusions. Thus, it is important to ensure that the performance impact of changes are not over-shadowed by other external factors.

Our approach. We propose a new approach, **Litmus**, to accurately assess the performance impact of changes in cellular networks and to tackle the over-shadowing impact of the external factors. The key idea is to compare the *relative performance* before and after the change at the study group (network elements where the change is implemented) and the control group (network elements without the change). Our algorithm learns a statistical dependency structure between the performance at the study and control groups before the change. The dependency structure captures how well one group can be used to forecast the other. It then uses the performance at the control group after the network change to construct the forecast values for the study group after the change. The hypothesis is that if there is no performance impact at the study group, then the dependency structure between the study and control groups should not change. Changes in the dependency structure are indicative of a performance impact induced by the network change.

We select the control group using a domain knowledge guided approach, with the control group having a similarity in certain attributes (*e.g.*, geographic location or software version) with the study group. The external factor has a higher likelihood of inducing a similar performance impact at both the study and the control groups (*e.g.*, weather changes will lead to a correlated degradation in performance across the study as well as the control groups).

The intuition behind study and control group analysis is based on three observations made using data collected from operational cellular networks. (i) Service performance at geographically nearby network elements is statistically correlated. (ii) External factors induce similar performance impact across multiple network elements. (iii) A performance impacting change at the study group creates a change in the relative performance between the study and the control group. By analyzing the relative performance between the study and the control groups, Litmus can effectively and accurately assess the performance impact of changes at the study group even in the presence of external factors.

Challenges. We need to address two key challenges: (i) Robust inference of the relative change in performance between the study and the control group - we want to avoid unrelated performance changes in a small number of control group elements to significantly and negatively impact the inference. (ii) Selecting network elements within the control group - an inaccurate selection would nullify the advantage of study and control group comparison. For example, when assessing the performance impact of changes at the study group elements in Northeastern regions of the United States that are influenced by foliage, it is important to select the control group elements from the regions influenced by foliage.

[1]We explicitly do not show any service performance numbers in the paper for proprietary reasons and risk of misuse by marketing. The observations and inferences still hold.

Our contributions.

- *Algorithm*: We proposed a new ***spatial regression algorithm*** in Litmus for doing a robust performance comparison of the study group and the control group especially when operationally it is impossible to guarantee a clean or eventless control group. Our algorithm is more effective than the study group only analysis [9, 18] and Difference in Differences (DiD) [21, 26].
- *Application*: We designed a ***domain knowledge guided control group selection*** that better captures the characteristics of external factors in cellular domain. Litmus is being successfully applied in operational cellular networks to assess the performance impact of changes. The output of Litmus is used in the decision making for a go or no-go for wide-scale deployment of changes. The "go or no-go" decision for the First Field Application (FFA) of the change is made by the Engineering and Operations teams based on the performance impacts (improvements or no degradations) induced by the network changes in the field. For example, service performance improvements observed at all of the FFA locations indicate that the network change would lead to an overall improvement of the quality of experience to the end-users and thus should be rolled out network-wide. Our experience has shown that external factors make the performance assessment in initial field tests difficult and comparisons between study and control group can help eliminate (or, reduce) the over-shadowing impact of external factors.

Paper organization. The rest of the paper is organized as follows. In Section 2, we provide background information on cellular service architecture and network changes followed by descriptions and motivating examples for why inferring the service performance impact of changes is hard specially in the presence of external factors. We present the design of Litmus in Section 3. In Section 4, we present a thorough evaluation of Litmus using data collected from operational cellular networks (known assessments as well as synthetic injection analysis) and demonstrate that it is more effective and accurate as compared to the study group only analysis as well as Difference in Differences [21, 26] approach in assessing the performance impact of changes in the presence of external factors. We share interesting case study findings in Section 5. We conclude in Section 6 with a discussion of future research opportunities in the area of change impact analysis.

2. BACKGROUND AND MOTIVATION

In this section, we first describe the cellular network architecture for Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS) and Long Term Evolution (LTE) (Section 2.1). In Section 2.2, we describe data sets relevant to the rest of the paper. In Section 2.3, we categorize the cellular network changes into high frequency changes to dynamically deal with changing network and traffic conditions and low frequency changes to optimize service performance over longer time-scales. Assessing the performance impact of changes has recently received a lot of attention and we summarize related work in Section 2.4. Finally, we use real-world data collected from operational cellular networks to demonstrate the technical challenges in assessing the impact of changes in cellular networks (Section 2.5).

2.1 Cellular Service Architecture

Fig. 2 shows the architecture for GSM, UMTS and LTE cellular networks. The cellular networks consist of three domains: User

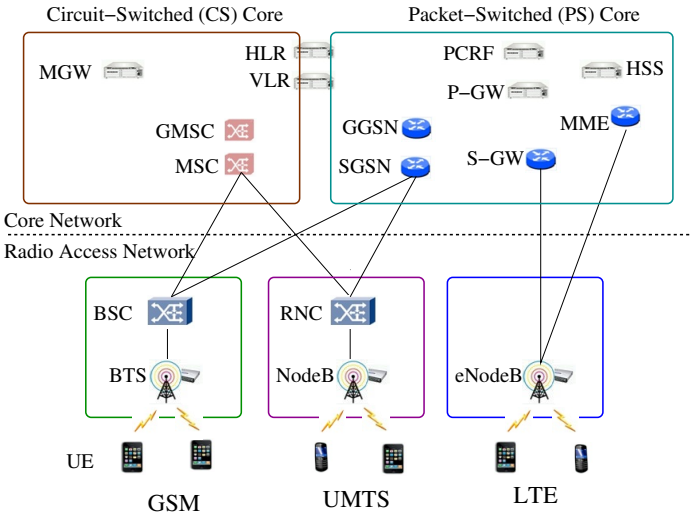

Figure 2: Cellular Network and Service Architecture for GSM, UMTS and LTE.

Equipment (UE), Radio Access Network (RAN) and Core Network (CN). The UE or mobile devices communicate using the air interface to the RAN network. They can either operate in Packet Switched (PS) mode where only data services can be accessed, Circuit Switched (CS) mode where only voice services can be accessed, or PS/CS mode where the UE is capable of simultaneously operating PS services and CS services.

Each RAN is subdivided into individual radio network systems (RNSs), where each RNS is controlled by a controller - Base Station Controller (BSC) for GSM, Radio Network Controller (RNC) for UMTS and evolved NodeB (eNodeB) for LTE. The main functions of the controllers are radio resource control, admission control, power control, handover control and broadcast signaling. In LTE, each controller (eNodeB) serves one or more cells, whereas, in GSM and UMTS, each controller connects to a set of towers (BTS and NodeB respectively), which each serve one or multiple cells. Generally, each cell tower has multiple transceivers or sectors. Each sector configurable by the frequency of operation, antenna tilt and downlink transmission power, determines the coverage area within a region. The UEs communicate via a serving sector and can hand-off to a neighboring sector (either on the same cell tower, or a different cell tower on the same radio access technology, or across technologies).

The core network is responsible for providing switching, routing and transit for user traffic. It also contains databases and network management functions. In GSM/UMTS, the core has separate elements for CS and PS. The main components for the CS core are the Mobile Service Switching Center (MSC) and Gateway Mobile Service Switching Center (GMSC). The MSC processes and tracks all incoming and outgoing voice calls. It is responsible for all call routing and billing functions as well as voice mail activity. The GMSC provides inter-connectivity with external circuit switched networks and cellular providers. The PS core consists of the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). The SGSN is responsible for transport and delivery of data packets to/from the UE, mobility management, authentication of the users and managing the logical link to the UE. The primary role of the GGSN is to route data from the Internet to the UE and vice versa.

The LTE core consists of the Policy Control and Charging Rules Function (PCRF), Home Subscriber Server (HSS), Serving Gateway (S-GW), Packet Data Network Gateway (P-GW), and the Mobility Management Entity (MME). All user IP packets are transferred through the S-GW. The MME is the control node which processes the signaling information between the UE and the core network. The P-GW is responsible for IP address allocation for the UE, as well as QoS enforcement and flow-based charging according to the rules from PCRF.

2.2 Data Sets

We now describe the data sets used in the rest of the paper. The cellular service provider collects a plethora of data from cell towers, radio network controllers, core switches, routers and servers. The data sets include performance measurements, call detail records (CDR), change management logs, and configuration snapshots.

Service performance measurements. Performance counters collected from individual network elements are used to compute aggregate service quality metrics such as voice and data accessibility, retainability, and data throughput. *Accessibility* is a measure of successful call attempts placed by users on the cellular network. An increase in call attempt failures leads to a lower accessibility. *Retainability* is a measure of successful call termination as issued by the user and not by the network. Network call terminations (also known as dropped calls) could happen due to several reasons such as radio interference, handover issues (intra or inter radio access technology), congestion, or radio link failure. Higher rates of dropped calls lead to lower retainability. Accessibility and retainability metrics are calculated separately for voice and data sessions. *Data throughput* captures the bits / bytes / packets delivered to users over the cellular network. We use these service quality metrics to assess the impact of network changes. They are also referred to as **K**ey **P**erformance **I**ndicators (KPI).

Network change management logs. The change management logs contain information about the changes and maintenance activities scheduled at the network elements. The cellular network changes can be either in the form of configuration changes (*e.g.*, antenna tilt modifications, radio link failure timers), software upgrades, topological changes (*e.g.*, re-homes of network equipment), hardware/firmware upgrades or traffic movements across data centers. They are often made to improve service performance, introduce new features, plan for big events (*e.g.*, Superbowl), or to reduce operational cost. We use the change information to determine when and where to perform the service performance assessments.

Network configuration. The configuration snapshots are collected on a daily basis and used to automatically infer the topological structure of the cellular network. The topological structure is used to identify (i) the causal impact scope of network changes (*e.g.*, neighboring cell towers), and (ii) the control group elements within geographical proximity of the study group (*e.g.*, cell towers sharing the common upstream radio network controllers).

2.3 Cellular Network Changes

Based on our observations of the types of changes in cellular networks, we categorize them into: (i) *high frequency* changes, and (ii) *low frequency* changes.

High frequency changes. Certain configuration parameters are dynamically changed based on network and traffic conditions. This is done by the cellular service providers to ensure rapid responses to changing network and traffic behaviors. For example, during a road accident or points of traffic congestion, the cell tower parameters can be dynamically adjusted to improve the service performance. Antenna tilt and power can be adjusted to balance traffic between cell towers. Decreasing power on a heavily loaded cell tower and increasing power on a lightly loaded neighboring cell tower can balance the load and lead to a better quality of experience to the end-users. Up-tilting the antenna increases the cover-

age area, while down-tilting the antenna reduces the coverage. We refer to these parameters as having a high frequency of changes. These high frequency change parameters are often controlled manually by the engineering teams and have *different values at different locations*. Recently there have been initiatives to deploy Self Optimizing Network (SON) [22] solutions that can automatically tune the network parameters in response to changing network and traffic conditions and improve service performance.

Low frequency changes. Configuration parameters that are changed on a longitudinal basis (*e.g.*, once in six months or a year) are referred to as *low frequency* changes. *Gold standard* configuration parameters have low frequency for their changes and are typically adapted during major software releases, or new feature roll-outs. An example of a low frequency gold standard parameter is a radio link failure recovery timer which would stay stationary over a long time and only changed with new service requirements or software loads. Typically, the recommendations for gold standard parameter changes come from the Planning and Engineering teams in collaboration with the equipment vendors. The gold standard parameters have the rule: *One value fits all locations* for easy management across the network.

2.4 Assessing Impacts of Changes

Once the changes are applied in the network, it is important to carefully identify their service performance impact. During the testing phase of FFA (First Field Application) changes, performance assessment plays a very important role in ensuring that the expected performance impacts (significant performance improvements and no degradations) are observed in the field. In order to ensure that the performance impacts of network changes are persistent, a longer time-scale (*e.g.*, 1-2 weeks) is typically selected for comparing the performance before and after the network change.

Related work. Mercury [20] compares the performance before and after the network change on a long-term basis (on the order of several days) and provides effective ways of conducting a network-wide assessment of changes. PRISM [19] however, focuses on a single network element under change and performs a near real-time performance assessment of network changes. Spectroscope [25] diagnoses performance changes by comparing two executions before and after the change. Reitblatt *et al.* [23] use per-packet and per-flow abstractions to ensure consistent behaviors during network updates. Canini *et al.* [5] use model checking for OpenFlow applications to capture subtle bugs during network updates. X-ray [4] uses differential performance analysis to compare the execution of two similar operations and explain why their performance differs.

None of the existing approaches focus on the problem of performance assessment of changes in the presence of external factors. They do not explicitly compare the performance at the study group with that of the control group. By focusing on study group only analysis, they could inaccurately infer the performance impact of changes in the presence of external factors. PCA [12, 13, 17, 24], SSA [10, 14] and compressive sensing [30] detect network-wide anomalies using unsupervised learning. However, they do not explicitly use the knowledge of study group versus control group network elements and could result in inaccurate inferences of the impact at the study group. For example, unsupervised learning would not be able to correctly identify a relative degradation at the study group compared to control when absolute improvements are observed across both the study and the control groups (Fig. 7(c) in Section 3.1). Using a supervised learning model is better tailored to the problem at hand. We believe that we are the *first* to use a supervised learning model and study group versus control group performance analysis in Litmus to accurately infer the impact of changes even in the presence of external factors.

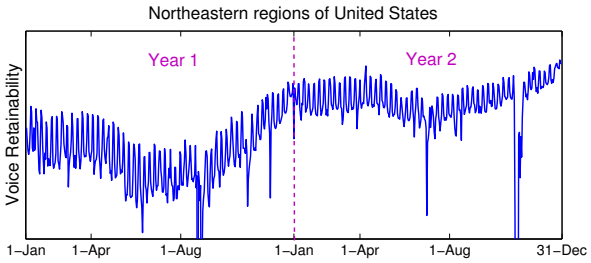

Figure 3: Seasonal patterns (due to foliage change) in Voice Retainability (daily aggregated) for UMTS cell towers at Northeastern region of United States using two years of operational network data.

Litmus comparison to A/B testing [15, 16]. A/B controlled experiments and testing (also known as split testing, control/treatment tests) are used in the web domains for making data-driven decisions. Web users are randomly exposed to one of the two variants of the experiments: control (A) and treatment or study (B). The user interactions are then recorded and key metrics are compared to analyze the impact of the treatment or study. In comparison to Litmus, the applications and domains are different. Study/control test and assessment are tied together in A/B testing, whereas in our problem, it is beyond our control when and where to execute the change. We explicitly focus on assessing the impact of executed changes. The control group might be subject to other events such as changes or unplanned outages. Thus, special consideration of robustness in regression and control group selection criteria is crucial in our context. It remains an interesting future challenge to design a change execution plan (under complex and massive operational constraints as well as foreseeable external factors such as weather, social events) for more effective impact assessment.

2.5 Why is Assessment Hard?

Assessing the impact of changes in operational cellular networks is hard because of the open nature of the system and the influence of external factors on service performance. In this section, using real-world data collected over two years from operational cellular networks, we will present illustrative and motivating examples that demonstrate the challenges of inferring the service performance impact of network changes in the presence of external factors such as seasonality, weather changes, traffic pattern changes and network events such as outages and other changes.

Seasonality. There is a significant influence of seasonality on the service performance metrics such as voice and data accessibility, retainability, and data throughput. Seasonality is observed at multiple time-scales: usage pattern of end users introduces a time-of-day effect (*e.g.*, high call volumes at peak hours of the day and low call volumes at night), and a weekly pattern (weekend versus weekday). Long-term seasonality at a yearly time-scale is seen in regions with extreme temperatures. Using two years worth of data collected from operational cellular networks (GSM, UMTS and LTE), we observe yearly seasonality across multiple service performance metrics. We show daily aggregated voice retainability for UMTS cell towers in the Northeastern regions of United States in Fig 3. There is a performance dip from April to August every year because of leaves budding, and a performance improvement from September to January because of leaves falling off the trees. Leaves create obstruction for radio signal propagation resulting in wireless signal fading. We see a better service performance when there are no

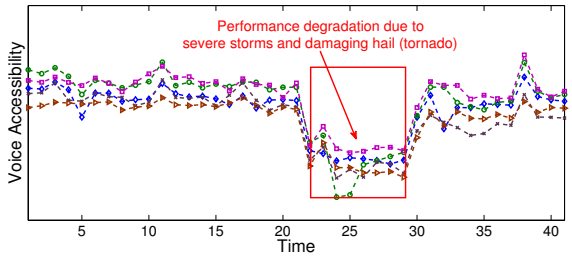

Figure 4: Performance degradation across multiple Radio Network Controllers (RNCs) due to severe storms and damaging hail during a tornado.

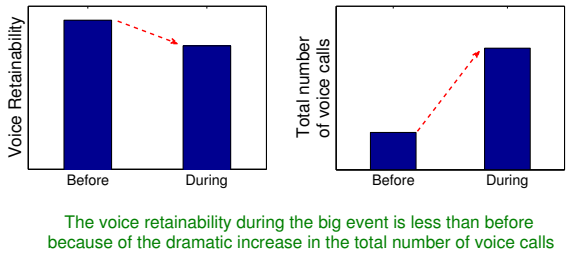

Figure 5: Traffic volumes and service performance for voice calls during a big event as compared to before.

leaves on the trees. We validate this seasonal trend (well recognized within the literature [6, 7]) using operational data. Performance assessment thus becomes challenging during time-intervals of impact due to foliage changes. Note that the overall increasing trends are likely due to the continuous improvements performed by the carrier on the network. Seasonal patterns however are not observed in the Southeastern region because of a lack of foliage change.

Weather changes. Different weather conditions such as rain, fog, snow and wind affect the voice as well as data sessions. Objects in the air such as rain, fog or snow have a negative impact on cellular service performance. Severe events such as massive storms, hail, tornadoes, or earthquakes can cause network outages (*e.g.*, tower failures, transport equipments out of service) and performance degradation for a longer time-interval. We collected weather data [1, 2] and compared it to the service performance data. We observed a service performance impact during days with continuous rainfall and high inches of rain - this is expected and we validated using real-world data. We further analyzed the severe weather events for 2012 collected from [1]. Examples of such events include hurricane Sandy, Midwest tornadoes, Rockies/Southwest Severe storms, and Texas tornadoes. For each event, we observed a performance impact across multiple cell towers. Fig. 4 shows the degradation in voice accessibility at multiple RNCs due to severe storms and damaging hail caused by tornadoes. Thus, if the time-interval of the storm had overlapped with a network change assessment, deriving conclusions on the performance impact of changes would become difficult.

Traffic pattern changes. It is well-known that the higher the traffic volume, the higher the likelihood of traffic being dropped (congestion scenarios or radio interference). We validate this using voice and data retainability performance metrics collected from several cell towers in the operational network. The traffic patterns can dramatically change during holidays or big events. Fig. 5 shows the voice retainability and total number of voice calls before and during a big event aggregated across all cell towers at the location of the event. As can be seen, the total number of voice calls

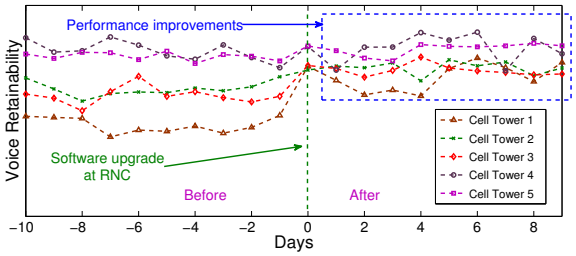

Figure 6: Performance improvements across a majority of cell towers due to software upgrade at upstream RNC.

dramatically increase during the event that induces an impact on the voice retainability (we make similar observations on data sessions). Thus, massive traffic changes can significantly influence the service performance metrics. It is therefore important to carefully account for traffic changes when assessing the service performance impact of network changes.

Network events. Events such as network changes, other maintenance activities or outages close in time for the FFA change under test can also significantly influence the performance assessment. Fig. 6 shows the performance improvements in voice retainability at multiple cell towers due to a software upgrade at an upstream RNC. This is indicative of a service performance improvement experienced by the users served by the upgraded RNC. From the perspective of impact analysis of the software upgrade at the RNC, this is a positive outcome. However, if configuration changes were made on a small number of cell towers (downstream to the upgraded RNC) around the same time, this would make performance assessment of the configuration changes at the cell towers difficult because of the overlapping software upgrade at the RNC. If the study group cell towers (those undergoing a configuration change) were analyzed in isolation, then the inference of performance improvement would be inaccurate because the root-cause is the software upgrade at the RNC and not the configuration change at the cell towers.

3. Litmus DESIGN

As described earlier, it is important to take overlapping external factors into account when assessing the performance impact of network changes. We propose a new approach, **Litmus**, that compares performance at the study group (network elements that have change implemented) with performance at the control group (network elements without the change). Our assumption is that external factors influence performance at both the study and control groups. By comparing the relative performance between the study and the control group before and after the network change, we improve the accuracy of the inference of the performance impact of changes at the study group by eliminating or reducing the over-shadowing impact of the external factors. We provide the intuition behind comparing study versus control group in Litmus in Section 3.1. We propose a new robust spatial regression algorithm to accurately compare the performance between the study and control groups before and after the network change (Section 3.2). We describe design choices for the selection of the control group in Section 3.3.

3.1 Intuition

We choose to compare the performance at the study group with that of a control group based on three observations in operational cellular networks: (i) geographically close network elements (*e.g.*, cell towers, radio network controllers) exhibit a high degree of spatial auto-correlation or statistical dependency in performance

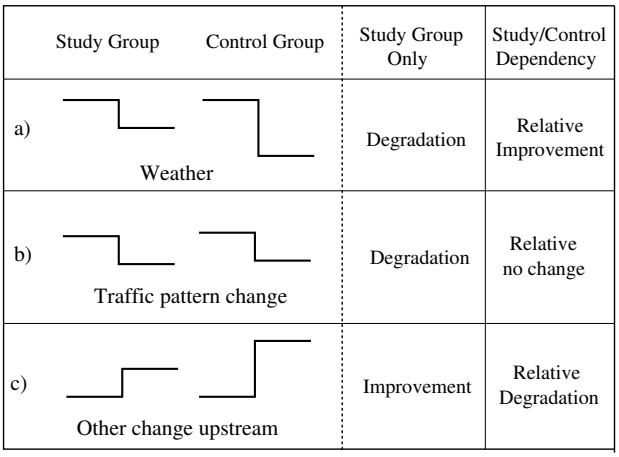

Figure 7: Illustrative examples of how study group only assessment can lead to incorrect inferences. Looking at the relative performance between the study and the control group helps in robust and accurate analysis of impacts.

because of the similarity in radio propagation, terrain (landscape, buildings), user population densities, and weather patterns, (ii) external factors have the potential to induce similar performance impacts (either improvements, or degradations) at multiple network elements, and (iii) a performance impacting network change at the study group induces a relative change in performance between the study and the control groups.

Fig. 7 provides a few examples of how comparing the relative performance between the study and the control group before and after the network change in Litmus provides accurate assessment of the impacts as compared to study group analysis only. As shown in Fig. 7(a), a weather event induces a performance degradation at both the study and the control groups but the network change at the study group results in a better relative performance at the study group compared to the control group. Fig. 7(b) shows the performance of a sudden traffic pattern change that resulted in a degradation at both the study and the control groups. Study group only analysis would infer a degradation, but comparing the relative performance between the study and the control groups demonstrates no relative change at the study group. Finally, in Fig. 7(c), a network change upstream of the study and control groups induces an improvement across both. But as can be seen, there is a relative degradation at the study group compared to the control.

Events such as service outages due to sudden weather changes, traffic pattern changes (*e.g.*due to road accidents), or network equipment failures are unexpected and not always easy to predict. Furthermore, if they occur around the time of the network change on the study group, then it becomes challenging to associate the performance impact to the change or other external factors. Even with predictable factors such as foliage, the change execution might be outside the control of Engineering and Operations teams because of market pressure and competition (*e.g.*, rolling out new service features such as Voice over LTE). These are exactly the scenarios where Litmus will be helpful and able to conduct a robust assessment of the change.

3.2 Study/Control Group Comparison

Given the time of the change at the study group, our goal is to compare the relative performance between the study and control group before and after the change. A simple straightforward approach is to compare statistical measures such as mean or median differences before and after the change. If $Y_b(j)$, $X_b(i)$ are the performance time-series before the change at study and control group elements j and i respectively, and similarly, $Y_a(j)$, $X_a(i)$ are the performance time-series after the change, then Difference in Differences (DiD) measure used in econometrics [21, 26] evaluates the metric for each pair of elements in study and control group:

$$d(j,i) = h(Y_a(j)) - h(Y_b(j)) - (h(X_a(i)) - h(X_b(i))) \quad (1)$$

$$j = 1..M \text{ and } i = 1..N$$

$$h(\cdot) = \text{mean}(\cdot) \text{ or } \text{median}(\cdot)$$

If there is no change in the relative performance between the study and the control group, the DiD measure should be near zero.

Why DiD does not work? The challenges with using the DiD approach are (i) poor selection of some of the control group elements can lead to poor forecasting, and (ii) sensitivity of the DiD measure to performance changes in a small set of network elements in the control group [3]. Such behaviors occur frequently in operational networks - time-of-day traffic patterns across locations, or changes in performance caused by other factors unrelated to the change under test. Thus, a few poorly selected network elements in the control group can create biased estimates in the DiD measure, which in turn can lead to incorrect inferences.

We will provide an illustrative example to demonstrate why DiD fails. Consider a cell tower that is serving a business location as the study and a cell tower that is serving a lake as one of the control group. Business locations are busy during the weekdays between 9 AM and 5 PM and lakes are visited by users typically either over the weekends or during evening times. The weekday/weekend traffic patterns would thus be quite different across the business and lake locations. Using the cell tower serving lake as one of the control group elements would definitely be a bad predictor. DiD does not take into account bad predictors. Thus, we need to consider robust models that can automatically take care of bad predictors and is data-driven. We now outline our robust spatial regression algorithm to detect changes in the relative performance between the study and control groups.

Robust spatial regression. Our algorithm learns a statistical dependency structure using a spatial regression between the performance at the study and the control groups before the change. The dependency structure captures how well one group can be used to forecast the other. Using control group as the predictors, it constructs the forecast values for the study group after the change. The forecast difference is computed by taking the difference between the observed study group performance and their forecast values. The forecast difference after the change is compared to the forecast difference before the change and statistically significant changes in the two are indicative of a change in relative performance between study and control group. If there is no performance impact of the change at the study group, then the forecast difference before and after the change are statistically indistinguishable.

We incorporate robustness into the learning process through uniform sampling on the control group and comparing the forecast difference results before and after the change for each sampling iteration. Multiple iterations of forecast difference comparison increase our confidence that a small number of performance changes in the control group do not significantly influence the outcome of the study and the control group analysis. Uniform sampling on the control group helps reduce the impact of poor predictors. If the study group comprises multiple network elements, we repeat the process for each network element in the study group and report the results individually. We also use voting to summarize across multiple elements in the study group. We now formally describe the algorithm.

X_b, X_a are the performance time-series matrices for the control group before and after the change. The columns represent time-series for each network element. $Y_b(j), Y_a(j)$ are the performance time-series for the j^{th} network element in the study group before and after the change. We uniformly sample (without replacement) k out of N control group elements ($k > \frac{N}{2}$). The sampling process for selecting the control group elements before and after the change is the same. Let us call the sampled matrices for the control group X_b^S and X_a^S. We learn the regression coefficients β using time-series before the change.

$$Y_b(j) = \beta X_b^S \tag{2}$$

Linear regression model is well-studied and nicely fits our purpose. Sparsity regularization is not desirable (*e.g.*, ridge [11], lasso [28] or l_1 [8, 29] regression) because we do not want changes in a very small number of control group elements after the change to significantly influence the forecast.

We compute the forecast values (Y_a^f) for the study group time-series after the change using regression coefficients β learned from equation (2) and sampled control group time-series X_a^S after the change.

$$Y_a^f(j) = \beta X_a^S \tag{3}$$

We aggregate the forecast values by computing the median across all the sampling steps: median($Y_a^f(j)$). The forecast difference time-series after the change is given by

$$Y_a^{diff}(j) = Y_a(j) - median(Y_a^f(j)) \tag{4}$$

Similarly, the forecast difference time-series before the change is

$$Y_b^{diff}(j) = Y_b(j) - median(Y_b^f(j)) \tag{5}$$

where, $Y_b^f(j) = \beta X_b^S$

We use robust rank-order tests [9, 18, 27] to statistically compare the forecast difference $Y_a^{diff}(j)$ with $Y_b^{diff}(j)$. If the forecast difference after the change is significantly greater than before the change, then we conclude that the study group has a relative increase as compared to the control group. On the other hand, if the forecast difference after the change is significantly lower than before the change, then the study group has a relative decrease as compared to the control group. No statistical change in the forecast difference indicates no relative change. We choose robust rank-order tests because they eliminate the undesirable impact of one-off outliers in the time-series and accurately identify change signatures such as level changes, and ramp-up/downs.

3.3 Control Group Selection

Since the performance impact evaluator knows if the impact is local or within some proximity, he/she selects the control group candidates outside of its impact scope. The robust spatial regression analysis can account for a small number of bad members in the control group, but if a majority (or all) of the control group elements are poorly selected, then it will lead to inaccurate forecast for the study group. An example of poor selection of the control group is cell towers in the Northeastern regions of United States when the study group is from the Southeastern regions (foliage changes at Northeastern regions will impact the performance forecast for the Southeastern regions). We propose two guidelines for the selection of the control group: (i) they are subject to the same external factors as the study group, and (ii) they share similar properties with the study group such as geographical proximity, configuration, or traffic patterns. The size of the control group plays an important role in the effectiveness of the forecast for the study group. If the size of the control group is too large, then it will be difficult to capture the similar impact of external factors and accommodate the similarity in properties with the study group. On the other hand, if it is too small, then we will loose the benefits of robust regression analysis for a few bad control group members. For Litmus evaluation as well as operational deployment, we intentionally set the control group to not be the whole network because external factors have limited impact (typical size = 10s-100s), keeping the scale manageable.

	OBSERVATION		
EXPECTATION	**Improvement**	**Degradation**	**No impact**
Improvement	TP	FN	FN
Degradation	FN	TP	FN
No impact	FP	FP	TN

Table 1: Methodology for labeling the algorithm outcome.

We design a domain knowledge guided control group selection. The knowledge about cellular network domain helps in creating attributes for selecting the control group. Litmus provides a wide range of attributes for control group selection:

1. Geographical distance using latitude, longitude and zip code
2. Topological structure of the cellular network
3. Configuration settings such software version, equipment model, or antenna parameters
4. Terrain using landscape, and buildings
5. Traffic patterns

These attributes were created through interactions with the domain experts responsible for the change and its impact assessment; hence, we term our approach domain knowledge guided.

We employ predicates to capture the dependency between the study and control group. Predicates can either be uni-variate (*e.g.*, cell towers within the same zip code), or multi-variate (*e.g.*, cell towers sharing the common set of upstream RNCs and upstream RNCs with same OS). The infrastructure for control group selection is flexible in terms of specifying predicates depending on the type of change.

4. Litmus EVALUATION

In this section, we present the evaluation of Litmus using data collected from operational cellular networks. Because of the lack of complete ground truth information, evaluation using real-world data is challenging. We address the issue by conducting the evaluation in two steps. First, using examples from real-world network changes and apriori known assessments by the Engineering and Operations teams, we quantify the accuracy of our robust spatial regression algorithm and show it outperforms Difference in Differences and analysis using the study group only. Second, we select performance time-series at cell towers and synthetically inject changes and case scenarios representative of seasonality, weather changes, holidays and network events. Synthetic injection provides a thorough and exhaustive evaluation of the algorithms across several scenarios. Thus, a combination of real-world and synthetic evaluation have boosted our confidence in the effectiveness of the study/control group analysis and robust spatial regression algorithm in Litmus.

4.1 Methodology

For evaluation using both known assessments and synthetic injection, we know the outcome of the assessment - either no performance impact, significant performance improvement or degradation. For each algorithm, we label the outcome as true positive

(TP), true negative (TN), false positive (FP) or false negative (FN). True positives are significant performance impacts accurately identified by the algorithm. True negatives are the instances that are correctly identified as not having a performance impact. False positives are incorrectly identified by the algorithm as having a performance impact when no impact is expected. False negatives are the outcomes the algorithm fails to capture as a significant performance impact. Table 1 shows how the algorithm outcomes are labeled. For example, if the expectation of the impact assessment is significant performance improvement, and the algorithm outputs significant improvement, then it is a true positive. However, if it either outputs significant performance degradation or no impact, then it would be tagged as a false negative. Given TP, TN, FP, FN, we can compute precision, recall, true negative rate and accuracy as follows: Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, True negative rate = $\frac{TN}{TN+FP}$, and Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$.

Using the above metrics, we compare the three algorithms: (i) study group only analysis, (ii) Difference in Differences, and (iii) Litmus robust spatial regression. For study group only analysis, we compare the study group performance time-series before the change with after the change. For Difference in Differences, the difference between study and control group is compared before and after the change. Finally, for the robust spatial regression algorithm, the forecast differences are compared before and after the change.

4.2 Evaluation using Known Assessments

In collaboration with Network Engineering and Operations teams, we collected a list of configuration changes in the production cellular networks including the expected outcomes and assessments of impacts performed by other teams. The Network Engineering and Operations teams have been carrying out the impact assessment process manually by visually inspecting the time-series before and after the change at the study group as well as the control group. This serves as the ground truth information about the performance impact of the changes at the study groups. For each study group, we identify network elements in the control group using the geographical measure (same zip code) for LTE and topological structure for GSM and UMTS (*e.g.*, NodeBs under the same RNC).

Table 2 summarizes the evaluation results for the three algorithms using known assessment of network changes in operational environments. The first column (Change Type) describes the type of the change and the second column (Location) indicates the type of network element it is applied on. The third column (Impact Expectation) indicates the Engineering and Operations teams' expectation of performance impact (*e.g.*, improvement ↑, degradation ↓, or no impact ↔). The fourth column (Impact Assessment) is the summary of the performance assessment *manually* conducted by the Engineering and Operations teams. The fifth column (External Factor) shows the presence of an external factor during the impact assessment. The sixth and seventh columns contain the number of elements in the study group and Key Performance Indicators (KPIs) for conducting the analysis. Columns 8-10 are the results for the three algorithms: study group only analysis, Difference in Differences and the robust spatial regression in Litmus. For each change, we identify the number of TP, TN, FP and FN and then summarize the precision, recall, true negative rate and accuracy. The control group was identified using the geographical proximity measure of same zip code/same technology for 109 out of 313 cases and topological structure/same technology for the remainder 204 out of 313 cases.

As can be seen from Table 2, the study group only analysis fails to take into account the expected changes in time-series induced by external factors such as seasonality, weather, holiday, other network changes, and foliage. This creates an incorrect inference of the performance impact in several cases. For example, the assessment of the SON (Self Optimizing Network) feature for automatic load balancing across cell towers (first row in Table 2) was expected to yield performance improvement because the load balancer automatically and dynamically adjusts the load between cell towers depending on the network and traffic behaviors, resulting in an improved quality of experience to the end-users. However, since the test in the field was being carried out in a region that had an impact due to foliage, the performance assessment of SON was negatively influenced by an external factor (in this case, foliage). Thus, by only focusing on the study group, one would inaccurately conclude a performance degradation because of the SON feature introduction. Since foliage impacts both the study and the control groups, the robust spatial regression algorithm in Litmus accurately detected the relative performance improvement. Difference in Differences accurately detected the relative impact in voice retainability and data throughput, but failed to correctly detect the impact in data retainability. The time-series patterns in data retainability made the inferences challenging and could only be correctly identified by Litmus.

Change Injection	Change Magnitude	Impact Expectation	Study Group Only Analysis	Study/Control Dependency Analysis
None		No	True Negative	True Negative
Study		Yes	True Positive	True Positive
Control		Yes	False Negative	True Positive
Study, Control	Same	No	False Positive	True Negative
Study, Control	Different	Yes	False Negative	True Positive

Table 3: Case scenarios for synthetic injection of changes in performance time-series.

The robust spatial regression algorithm in Litmus performed extremely well on all of the 313 case scenarios, yielding 100% accuracy. Difference in Differences (DiD) had zero false positives giving 100% precision, but because of false negatives (missed detections of expected performance impact), its accuracy is 84.66% and recall is 79.49%. DiD's performance is similar to the robust spatial regression algorithm in Litmus except for a few cases that caused false negatives. Even though such cases are rare, they do exist in operational networks and Litmus provides an even more robust solution. With the study group only analysis, the metric values were the lowest (41.53% accuracy, 56.09% precision, 61.14% recall and 0.98% true negative rate) amongst the three, because of the high false positives and false negatives created by external factors.

4.3 Evaluation using Synthetic Injection

Having evaluated the algorithms using known assessment of changes in production networks, we now turn to thoroughly evaluating them across different case scenarios and time-series characteristics of the control group. We select study group performance time-series for voice and data accessibility, and retainability for UMTS cell towers from four geographically diverse regions (Northeastern, Southeastern, Western and Southwestern) in the United States. Cell towers in the Northeastern regions have yearly seasonality due to foliage, whereas other towers have been impacted by regional network changes. We also picked a few study groups which had no performance impact. The control group network elements are selected using the topological structure.

For each study group and a list of network elements in the control group, we synthetically injected changes such as level shifts in the performance time-series. Table 3 summarizes the case scenarios for the time-series patterns and the output expectations of the study group only analysis versus study/control dependency analysis. We

Change Type	Location	Impact Expectation	Impact Assessment	External factor	Number of study group elements (a)	Number of KPIs (b)	Study Group Only Analysis	Difference in Differences	Litmus Robust Spatial Regression
SON load balancing	RNC	Data & Voice Retainability ↑ & data throughput ↔	↑	Foliage	18	3	36 FN, 18 FP	18 TP, 18 TN, 18 FN	36 TP, 18 TN
Radio link failure timer	RNC	Voice Retainability ↑	↑		3	1	3 TP	3 TP	3 TP
Power	NodeB	Data Throughput ↑	↔		1	1	1 TN	1 TN	1 TN
Radio link	NodeB	Voice Retainability ↑	↔	Other change	25	1	25 FP	25 TN	25 TN
Power	RNC	Data Retainability ↔	↔	Other change	16	2	32 FP	32 TP	32 TP
Update new UE types	MSC	Voice Retainability ↑	↔	Seasonality	3	1	3 FP	3 TN	3 TN
Data parameter	RNC	Data & Voice Retainability ↑	↑		2	3	6 TP	4 TP, 2 FN	6 TP
Limit max power	RNC	Data Throughput ↑	↔	Holiday	3	1	3 FP	3 TN	3 TN
Access threshold	RNC	Voice Retainability ↑	↑		1	1	1 TP	1 TP	1 TP
Time to trigger	eNodeB	Data Accessibility ↑	↑		1	1	1 TP	1 TP	1 TP
Radio link	BSC	Voice Retainability ↑	↑		1	1	1 FN	1 TP	1 TP
Timer changes	RNC	Voice Retainability ↑	↑		5	5	20 FP, 5 TP	20 TN, 5 TP	20 TN, 5 TP
State transition features	RNC	Voice Retainability ↔	↓		1	1	1 TP	1 TP	1 TP
SON neighbor discovery & load balancing	RNC	Data & Voice Retainability ↑	↑	Weather	2	4	8 FN	8 TP	8 TP
Reduce downlink interference	eNodeB	Data Accessibility, Retainability & Throughput ↑	↑		30	3	90 TP	90 TP	90 TP
Handover	RNC	Data & Voice Retainability ↑	↑		19	2	28 FN, 10 TP	28 FN, 10 TP	38 TP
Inter-system Handover	RNC	Voice Retainability ↑	↑		3	1	3 TP	3 TP	3 TP
Software	eNodeB	Data Retainability ↑	↑		9	1	9 TP	9 TP	9 TP
Software	eNodeB	Radio bearer ↔	↓		9	1	9 FN	9 TN	9 TN
Summary					Total of 313 cases ($\sum a * b$)		129 TP, 1 TN, 101 FP, 82 FN	186 TP, 79 TN, 0 FP, 48 FN	234 TP, 79 TN, 0 FP, 0 FN
Precision = $\frac{TP}{TP+FP}$							56.09 %	100.00 %	100.00 %
Recall = $\frac{TP}{TP+FN}$							61.14 %	79.49%	100.00 %
True negative rate = $\frac{TN}{TN+FP}$							0.98 %	100.00 %	100.00 %
Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$							41.53 %	84.66 %	100.00 %

Table 2: Evaluation results using known assessments of network changes. The control group was appropriately chosen (in 10s-100s, not shown). Symbols used for performance improvement, degradation and no impact are ↑, ↓ and ↔, respectively.

confirm a strong statistical dependency in the performance time-series between the study and the control group and thus a synthetic change injected in either the study or control group would create a change in the dependency structure. For changes injected in both study and control group, the magnitude of the injected change determines the outcome of impact. For the same magnitude, we expect no impact and for different magnitude changes, we do expect performance impact. The changes injected are representative of the performance impact due to external factors.

The third column in Table 3 captures the impact expectation and the last two columns indicate the correct/incorrect inferences of the algorithms. We also introduced a noise component (level change) in a small number of control group elements to make the dependency learning challenging. This will help us compare the Difference in Differences approach with our robust spatial regression algorithm in Litmus. We compare 14 days before the change with 14 days after the change to determine the outcome of the assessment.

Table 4 summarizes the results for the three algorithms. A total of 8010 case scenarios have been evaluated (TP+TN+FP +FN). The study group only analysis lacks the ability to account for the performance impact due to external factors and thus leads to low precision, recall, true negative rate and accuracy compared to the other two algorithms that compare study group with control. The robust spatial regression algorithm outperforms Difference in Differences because of the robustness to deal with contamination of the control group. Difference in Differences have an accuracy of 75.43% which is outperformed by our algorithm in Litmus (accuracy of 82.35%). Our algorithm has a slightly higher number of false positives compared to Difference in Differences but maintains a low false negative (low misses). This is reflected in the true negative rates (41.19% for Difference in Differences versus 37.21% in Litmus). However, Litmus has a better recall (97.47%) than Difference in Differences (86.90%).

	Study Group Only Analysis	Difference in Differences	Litmus Robust Spatial Regression
True positive	4454	5214	5848
True negative	75	828	748
False positive	1935	1182	1262
False negative	1546	786	152
Precision	69.71 %	81.52 %	82.25 %
Recall	74.23 %	86.90 %	97.47 %
True negative rate	3.73 %	41.19 %	37.21 %
Accuracy	56.54 %	75.43 %	82.35 %

Table 4: Evaluation results using synthetic injection.

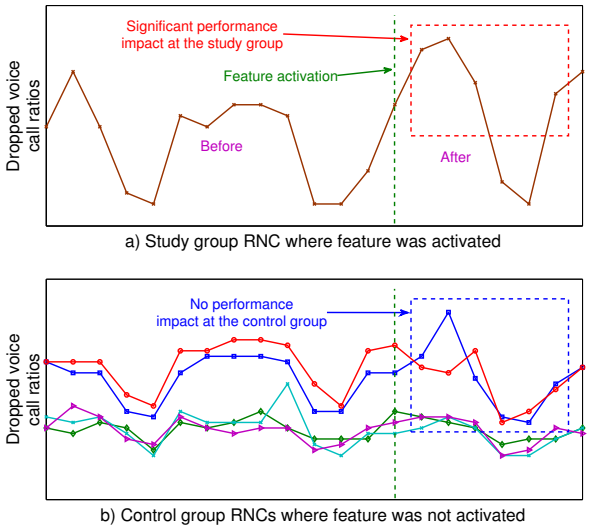

Figure 8: Significant increase in the dropped voice call ratios due to feature activation at a Radio Network Controller (RNC). Study/control group analysis in Litmus accurately captured the impact.

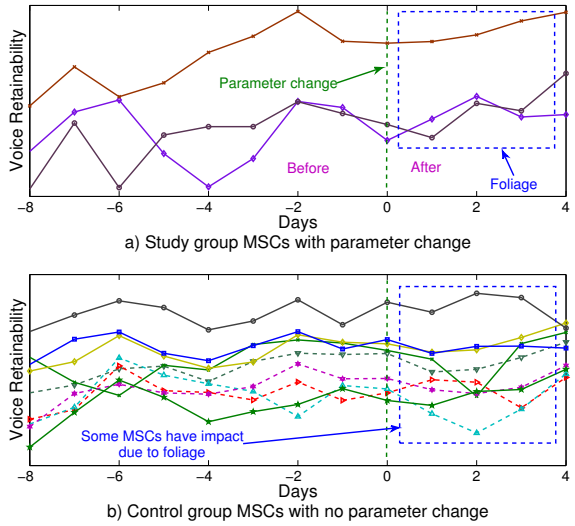

Figure 9: Foliage induces improvement in voice retainability at both the study and control group MSCs. The parameter change at the study group MSCs is thus not the cause for the performance improvement. This would be a false positive with the study group only analysis.

4.4 Summary

Using real-world known assessment of network changes and synthetic injection of changes, we have thoroughly evaluated the three algorithms across a wide range of case scenarios. We make two conclusions: (i) Comparing the study group time-series with the control group is extremely important in accurately inferring the performance impact of changes. Study group only analysis can lead to incorrect inferences due to performance impacts induced by external factors such as seasonality (foliage), weather changes (storms), traffic pattern changes (holidays) or other network events (outages or changes). (ii) Using the similarity in performance time-series between the study and the control group (*e.g.*, geographically close cell towers or those sharing the same upstream network elements), we can detect if a network change induces a change in the dependency structure between the study and the control groups. Our approach is thus robust to external overlapping factors and can accurately assess the impact of changes at the study group.

5. OPERATIONAL EXPERIENCES

We now present our case study experiences in applying Litmus on operational cellular networks (GSM, UMTS and LTE). Litmus is being successfully applied to assess the performance impact of network changes and provide inputs to decisions regarding go or no-go for wide-scale deployment. The case studies highlight the advantage of using the study group and control group dependency analysis to accurately determine the performance impact of changes. It is common operational practice to confirm performance impacts over multiple time-intervals before a decision is made for a wide-scale roll-out. The assessment time-scales are multiple days, typically 1-2 weeks, and our algorithm finishes in a few minutes, sufficient for our application. In comparison to the manual assessment conducted previously by the Network Engineering and Operations teams, Litmus provides an automated, scalable, easy-to-use and effective solution.

The first case study demonstrates that our approach does not miss performance impact captured by the study group only analysis techniques. Case studies 2-4 show that our approach accurately assesses the performance impact of changes even when there is an impact due to external factors such as foliage, weather changes, and holidays.

5.1 Impact of feature activation at RNC to reduce start-up times for data

In our first case study, we demonstrate that our approach can accurately identify the performance impact of a feature activation in the field. The Engineering teams were testing a new feature release at a RNC (Radio Network Controller) that was aimed at reducing the start-up times for data sessions in the UMTS network. They carefully analyzed multiple service performance indicators before and after the feature activation.

Using our approach, we compared the service performance metrics at the RNC that had the feature activated (study group) with other RNCs in the region (control group) that did not have the feature activated. We found that there was an unexpected and persistent impact in voice retainability at the study group. This was a confirmation of an earlier finding - a dropped voice call issue was found to be in the core network. The feature has been rolled back and scheduled to be tested with the new software release.

Fig. 8(a) shows the time-series for the dropped voice call ratios at the study group RNC where the feature was activated. As we can see, there is a subtle statistical change in the time-series after the feature was activated. Fig. 8(b) shows the time-series for other RNCs in the region where the feature was not activated. There was no change in the time-series behavior at the control group. Using our approach, we determine that the significant increase is indeed caused by the new feature - the forecast time-series for the study group RNC was statistically different than the observed time-series.

5.2 Impact of configuration changes at MSC influenced by foliage

In our second case study, we focus on the configuration changes at the MSC (Mobile Services Switching Center) in the UMTS network. The changes had been applied in Fall and the expectation of the Engineering teams was performance improvement in voice retainability metrics. Since the changes were applied at MSCs in the Northeastern regions of United States, foliage played a significant

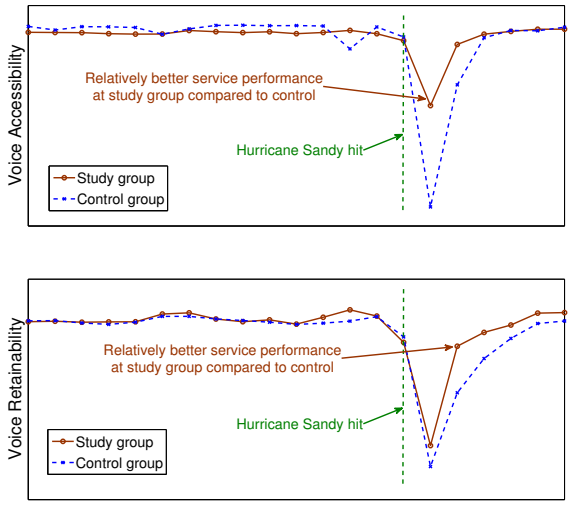

Figure 10: Litmus accurately captured relative improvement in service performance during hurricane Sandy at the study group cell towers with SON capabilities compared to the control group without SON capabilities.

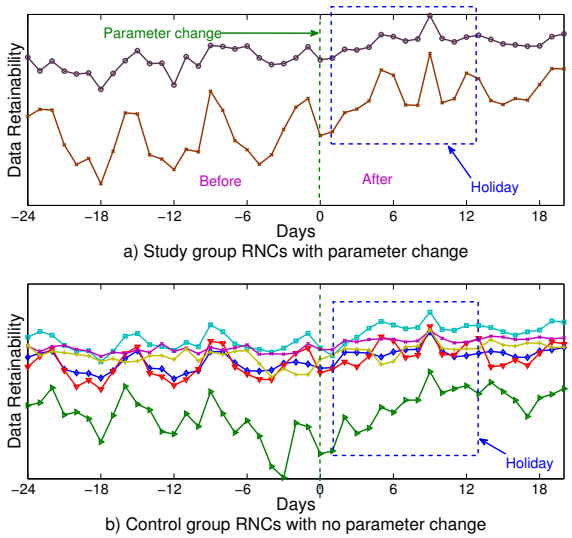

Figure 11: Significant increase in data retainability at study group RNCs is due to holidays. The impact is seen at both the study and the control groups. This would be a false positive with the study group only analysis.

role in influencing the assessment of the performance impact. Because of the pressure to deliver the new service features, the service provider could not wait for the foliage effect to disappear. Thus, the impact assessment had to be conducted during the Fall and with the impact of foliage on service performance. Fig. 9(a) shows the voice retainability at the study group MSCs with the configuration change. If we use the study group only analysis, we would infer that there are performance improvements because of the change. However, foliage played a role in improving the voice retainability metrics at multiple MSCs (both study and control group) across the Northeastern regions. Fig. 9(b) shows the voice retainability at the control group MSCs without the configuration change. As can be seen, only a few of the control group MSCs have improvement due to foliage. This is due to the geographical location of the MSCs and different intensities of foliage.

The Engineering teams used Litmus to carefully assess the performance impact of configuration changes at the MSCs and attributed foliage to the performance improvement. Litmus showed that there was no change in the relative performance (voice retainability) between the study group and control group MSCs. Thus, the robust spatial regression algorithm in Litmus played an important role in the accurate inferences of the performance impact. The Engineering teams further used Litmus to confirm that there were no unexpected performance degradations because of the configuration change and decided to keep the configuration change at the study group MSCs.

5.3 Impact of SON parameter changes at cell towers during hurricane Sandy

We now use Litmus to assess the performance impact of SON (Self Optimizing Networks) parameter changes at cell towers during hurricane Sandy that affected several areas in the Northeastern region of United States. The cellular service provider had deployed SON capabilities for automatically discovering cell tower neighbors (newly added as well as deleted due to failures) and load balancing across cell towers. These features were deployed well before Sandy, and not in anticipation of the hurricane. The hurricane Sandy served as a strong test for the deployed SON capabilities in the operational environments. Not all towers were SON enabled because the feature testing was on-going. During congestion or outage scenarios, SON automatically tunes the configuration parameters at the cell towers and optimizes the service performance. We had been tasked with assessing how good or bad did SON perform during hurricane Sandy.

If we use study group only analysis, clearly several of the service KPIs noticed an absolute degradation because of the cell tower outages. This is expected because of the significant impact of the hurricane. However, in order to assess the effectiveness of SON, we need to compare the study group (SON-enabled cell towers) with the control group (cell towers that did not have SON features activated). Using the robust spatial regression algorithm in Litmus to compare the performance of the study group with the control group, we observed significant relative improvements in service performance (voice and data accessibility as well as retainability) at the study group. Fig. 10 shows that the service performance metrics for the study group are relatively better than for the control group.

We confirmed with the Engineering teams and SON equipment vendor that the automatic neighbor discovery and load balancing features indeed contributed to better relative performance. Litmus thus concluded that even though there was an absolute degradation across the cell towers because of the hurricane, SON enabled cell towers did a good job as compared to those that were not enabled. SON dynamically improved the service performance across multiple cell towers. This further motivated a roll out of the SON features across the entire network. This case study demonstrated the effectiveness of comparing the study group with the control group when assessing the performance impact of changes at the study group in the presence of major weather events.

5.4 Impact of parameter changes at RNC to improve cell change success rates

In our final case study, we show that our approach is robust to the impact due to holidays. A parameter change to improve the cell change success rates was being tested at a few RNCs. Fig. 11(a) shows the data retainability at the study group RNCs. Using the study group only analysis, we found that there are significant in-

creases in data retainability after the change. This actually turned out to be a false positive because the holiday season was significantly influencing the performance indicators at all RNCs in the region (both study and control group). If we had relied on the study group only analysis, then we would have incorrectly inferred performance improvements and then recommended to roll out the parameter change across all the RNCs in the network. Fig. 11(b) shows that the time-series at the control group RNCs where the parameter was not changed, also observe a significant increase. It is thus important to compare the study group with the control group and assess if there is any relative improvement.

Litmus accurately captured this behavior in the forecast using our robust spatial regression algorithm and labeled the parameter change as having no impact. We confirmed this result with the Engineering teams - their response was that this finding of Litmus was very good because they have to be confident about the performance assessment of a change before they roll it out across the entire network. The decision of the Engineering teams was not to roll out the parameter change at other RNCs because the change did not contribute to a performance improvement at the study group RNCs.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we focused on the problem of assessing the performance impacts of changes in cellular networks in the presence of external factors such as seasonality (due to foliage), weather changes (storms, hurricanes), traffic pattern changes (*e.g.*, during big events, holidays), and network events such as outages or other changes. We proposed a new approach, **Litmus**, to tackle the overshadowing impact of the external factors. Litmus uses a robust spatial regression algorithm to detect changes in the relative change in performance between the study group and the control group before and after the network change. It thus accurately assesses the performance impacts of network changes at the study group even in the presence of impacts due to external factors. We systematically and thoroughly evaluated Litmus using real-world data collected from operational cellular networks as well as synthetic injection of changes. The robust spatial regression algorithm outperforms the study group only analysis and Difference in Differences. Our operational experiences have demonstrated the effectiveness of our approach. Litmus is now being successfully used as input to decisions regarding go or no-go for wide-scale deployment of changes in production cellular networks.

The design principles in Litmus are applicable to assessing the performance impacts of changes in other domains such as multi-tier cloud services, data center applications. These domains have complex dependencies across multiple components and network changes can be influenced by impacts due to many factors. In the future, we plan to apply Litmus in cloud and data center services. It is also interesting to expand the change impact assessment across different types of devices such as Apple iPad, Nokia Lumia, or Samsung Galaxy. The large number of combinations of device attributes (type, model, and version), different baseline and traffic behaviors across devices depending on popularity and usage, and dependency of service performance on network events would make the problem challenging. We plan to extend Litmus to monitor the impact of network changes on device performance and the impact of device upgrades on service and network performance.

Acknowledgment

We thank Aman Shaikh, our shepherd Ruben Merz and the CoNEXT anonymous reviewers for their insightful feedback on the paper. We are grateful to Giritharan Rana, Jia Li, Spencer Seidel, and Carsten Lund for their invaluable help on setting up the data feeds and reporting infrastructure for easy and flexible use on an ongoing basis. We strongly appreciate the collaboration and continuous support from the Network Engineering and Operations teams in the application of Litmus, regular feedback to improve its usability in the field and case-study analysis. We thank Mukta Mahimkar for recommending the name Litmus.

7. REFERENCES

[1] National climatic data center. `http://www.ncdc.noaa.gov`.
[2] Weather forecast and reports. `http://www.wunderground.com`.
[3] A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 2005.
[4] M. Attariyan, M. Chow, and J. Flinn. X-ray: Automating root-cause diagnosis of performance anomalies in production software. In *USENIX OSDI*, 2012.
[5] M. Canini, D. Venzano, P. Peresini, D. Kostic, and J. Rexford. A NICE way to test openflow applications. In *USENIX NSDI*, 2012.
[6] T. H. Chua, I. J. Wassell, and T. A. Rahman. Wind-induced slow fading in foliated fixed wireless links. In *IEEE VTC*, 2012.
[7] B. L. Cragin. Prediction of seasonal trends in cellular dropped call probability. In *Electro/Information Technology*, 2006.
[8] D.L.Donoho. For most large underdetermined systems of equations, the minimal l1-norm near solution approximates the sparsest near-solution. In *http://www-stat.stanford.edu/ donoho/Reports/*, 2004.
[9] N. Feltovich. Nonparametric tests of differences in medians: Comparison of the wilcoxonmann-whitney and robust rank-order tests. *Experimental Economics*, 2003.
[10] H. Hassani, S. Heravi, and A. Zhigljavsky. Forecasting european industrial production with singular spectrum analysis. *International Journal of Forecasting*, 2009.
[11] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.
[12] L. Huang, X. Nguyen, M. Garofalakis, A. Joseph, M. Jordan, and N. Taft. In-network PCA and anomaly detection. In *NIPS*, 2006.
[13] Y. Huang, N. Feamster, A. Lakhina, and J. J. Xu. Diagnosing network disruptions with network-wide analysis. In *ACM SIGMETRICS*, 2007.
[14] T. Ide and K. Tsuda. Change-point detection using krylov subspace learning. In *SIAM International Conference on Data Mining*, 2007.
[15] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: five puzzling outcomes explained. In *ACM KDD*, 2012.
[16] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *ACM KDD*, 2007.
[17] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM*, 2004.
[18] J. R. Lanzante. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station. *International Journal of Climatology*, 1996.
[19] A. Mahimkar, Z. Ge, J. Wang, J. Yates, Y. Zhang, J. Emmons, B. Huntley, and M. Stockert. Rapid detection of maintenance induced changes in service performance. In *ACM CoNEXT*, 2011.
[20] A. Mahimkar, H. H. Song, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and J. Emmons. Detecting the performance impact of upgrades in large operational networks. In *ACM SIGCOMM*, 2010.
[21] B. Meyer. Natural and quasi-experiments in economics. *Business and Economic Statistics*, 1995.
[22] J. Ramiro and K. Hamied. Self-organizing networks (SON): Self-planning, Self-optimization and Self-healing for GSM, UMTS and LTE.
[23] M. Reitblatt, N. Foster, J. Rexford, C. Schlesinger, and D. Walker. Abstractions for network update. In *ACM SIGCOMM*, 2012.
[24] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar. ANTIDOTE: understanding and defending against poisoning of anomaly detectors. In *ACM IMC*, 2009.
[25] R. R. Sambasivan, A. X. Zheng, M. D. Rosa, E. Krevat, S. Whitman, M. Stroucken, W. Wang, L. Xu, and G. R. Ganger. Diagnosing performance changes by comparing request flows. In *USENIX NSDI*, 2011.
[26] W. Shadish, T. Cook, and D. Campbell. Experimental and quasi-experimental designs for generalized causal inference. *Houghton Mifflin Co., Boston, MA*, 2002.
[27] S. Siegel and N. J. J. Castellan. Nonparametric statistics for the behavioral sciences. *New York: McGraw-Hilli*, 1998.
[28] R. Tibshrirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, 1996.
[29] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *ACM IMC*, 2005.
[30] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu. Spatio-temporal compressive sensing and internet traffic matrices. In *ACM SIGCOMM*, 2009.