Introduction

This report summarizes our validation findings for our Syraq model. In this report we first review our experimental setup and then test for the existence of any statistical error in our forecasts. Then in order to measure the characteristics of that error, we propose three null hypothesis models that can be used to judge our forecast error, test those models against six different out-of-sample cut points, and then show preliminary results from three short ablation studies. For context, please see our “Making Sense of Syraq” report.¹

Experimental Setup

During this project we used our Badlands Syraq model of the geographic area comprising Syria and Iraq to run a “retrodictive” experiment, attempting to test our model’s out-of-sample accuracy of territory control between October 2015 and March 2016. For our ground truth data, we used a dataset developed by Lustick Consulting that codes territory control at the

district-week level for 100 weeks (April 2014 to March 2016) and 162 districts (N=16,200, Territories include Assad, Rebels, ISIS, Iraq, and Kurds). We then split the data into two samples on the date where the 80th percentile of transitions occurs in our full dataset, which leaves an out-of-sample dataset that is 25 weeks long (N=4050). We ran 200 runs, using the first 80% of the data to steer the model in-sample.

Testing for Existence of Error

Below we test for the existence of error in our out-of-sample forecasts. We first test this hypothesis by using a chi-squared test during the final timestep of our model output to determine if and how well our 200 runs match with the ground truth data. The second test measures whether our model predicted the correct number of transitions for each district. The tests find very strong evidence that our model results deviate from zero error. Since our model’s target is the range of plausible futures that could occur, not the one actual future that will occur, and since model results are here measured against the one actual future that happened to occur, this is not surprising.

Method 1 - Chi Square

Pearson’s Chi Square test is a classic method of testing the independence of categorical datasets. It is calculated as the sum of the squared differences between different categories, and as a result significantly punishes major errors. The test requires that the observations going into the distribution are independent of each other, and the normalization of the differences requires that the expected value of any cell can not be zero. The first requirement suggests that we limit our analysis to one timestep per test, and the second demands that we do not use the standard ‘contingency table’ format as that would require zeroes in some of the expected results cells.

<table>
<thead>
<tr>
<th>Time</th>
<th>chi</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
<td>161</td>
<td>5.09883885135926e-238</td>
</tr>
</tbody>
</table>

We measure the total chi-squared distance per district between a perfect model (all 200 runs are correct) and our model results. The table above shows the chi-squared distance, degrees of freedom, and p-value for the final timestep of our model results. We find that there is a vanishingly small probability that the model and the ground-truth distributions are the same, with a likelihood around $10^{-238}$.

Below we measure the chi-squared distance for each timestep in the data (see Figure 1). The horizontal line shows the chi-squared distance necessary for a p-value above .05, therefore allowing us to not reject the null hypothesis. Besides the first timestep (which features five one-timestep transitions that then reverts), our model error tends to increase over time as our model drifts away from the real world.
Figure 1 - The distance between the model and our ground-truth data has been calculated for every out-of-sample timestep, and graphed above. As expected, the error increases as the model and the real world diverge and evolve beyond the constrained in-sample period. Timestep 96 represents October 4th, 2015 and timestep 120 represents March 18th, 2016.

Method 2 - T-Test of District Transitions

Beyond accuracy of each prediction, we can test other aspects of the model behavior. For the second test, we examine how many control transitions each district experienced, both in the model and in ground truth. We would expect that each district will have the same number of transitions in both datasets if our model did not exhibit error. A t-test assesses whether that statement is true on average.

For each district in each run in our data, we measure the number of transitions during the out-of-sample period. We then subtract that value from the number of transitions in our ground truth data. If a model run performs perfectly then the difference will be zero, it will be negative if our model overpredicts, and positive if our model underpredicts. Figure 2 below shows this data in a histogram. Below that we show the t-test results, making clear that it is very unlikely that our transition differences have a mean of zero. The t-test results show that our model tends to be short by half a transition per district, a result which is visible in Figure 2.²

² It should be noted that this test could also be performed using the absolute value of transition differences, but since we found no evidence of a match in this initial analysis the absolute value case would by definition produce an even more extreme result.
Measuring Direction of Error

Method 3 - Brier Score

The Brier score was originally developed for use in the field of meteorology. Unlike Chi-square and T-Tests, Brier scoring has a long history of being used on time-correlated categorical data. Defined as the mean-squared error, the Brier score compares the difference between the probabilistic prediction (between 0 and 1) and whether or not the categorical event occurred (0 or 1) with lower scores being better. The Brier score is a “proper scoring method”, meaning that it generally rewards honesty about uncertainty in forecasts. Since both Chi-square and the Brier Score are built on squared error, the shape of our model's Brier score over time in Figure 3 below and Chi-squared distance over time in Figure 1 are identical. The main difference is that the Brier score is agnostic to sample size.

Figure 2 - Distribution of differences between the model and ground-truth transitions. There is a very strong clustering around zero, but noticeably longer negative tail.

Figure 3 - This graph shows the Brier score over time for our Syraq ABM forecasts of territory control. The forecast error tends to increase over time.

---

Null Hypothesis Models

Direct comparison to the real world is the ultimate test of our model. However, given the complexity of our subject the mere presence of error is not enough to discount the results or the validity of the model. This is because there are no other existing strategies that can forecast territory control by district-week available, so we need to test whether our model is the best imperfect tool for the job. Below we have identified three possible “null hypothesis” models that can be used to test how wrong our model is compared to a baseline. Although there are an infinite number of potential models to compare, we have chosen three we think are reasonable. It should be noted, of course, that while our model produces value added in the form of detailed trajectories accounting for how outcomes predicted were accomplished, the three models used to produce the null hypotheses are strictly correlative, offering no claims about causal mechanisms. (One might also think about these models as different types of “dart-throwing chimps” that have their own systematic method for making forecasts in a “random” way.)

Carry Forward

The carry forward model simply assumes each district will remain in the territory it was in the previous timestep until the end of the dataset. This model will always miss transitions since it depends entirely on the most recent state of the model.

Modal

The modal model uses each district’s history within the in-sample dataset, and the most common zone for each district is then the forecast for all out-of-sample timesteps. This model treats all historical data as equally relevant to future forecasts.

Static Frequency

A slight refinement over the modal hypothesis, the static frequency model calculates a probability of a district being in each zone using the in-sample data. This probability is used as the forecast. While the static frequency forecast is almost never perfectly right, it tends to avoid being 100% wrong as well. This model also treats all historical data as equally relevant to future forecasts.

Figure 4 - Left: Brier scores for the ABM model and the three ‘null hypothesis’ models for the last week in our dataset. Right: Brier scores over time in all four models.
In Figure 4 above, we can see that our model does fairly well in the match-up against the three null hypothesis models, doing noticeably better than the modal and static probability models, but losing to the carry forward model. Few of the differences during the last week of our data are statistically different from each other, indicating a relatively static dataset with few transitions.

![Brier Scores by Zone](image)

**Figure 5 - Brier Score by zone and by model.**

Figure 5 above shows the Brier score for each model broken down by zone. Interestingly, the error is not distributed evenly across zones or models. The modal and static probability models unsurprisingly share a similar distribution of error, with the chaotic and opportunistic ISIS at the top. Al Nusra has the least error because the zone is so small and rare. The error for the carry forward model is defined by the number of transitions, from which we conclude that Iraq is a remarkably stable zone. Finally, our the ABM’s error is concentrated in the Islamic State territory and Assad’s Syria. Our ABM predictably has greater difficulty forecasting the chaotic and shifting Syrian conflict than the more stable Iraqi side.

![Brier Scores by District](image)

**Figure 6: A similar breakdown of sources of error by model, but by district instead of zone**

---

4 We found that although the average ABM Brier score is not statistically different from the scores for the carry forward and static models, the t-test between the ABM and modal null hypothesis Brier scores reports a p-value of 0.027 indicating a likely difference between the two.
In Figure 6 above, we show the Brier score breakdown by district in districts where there is ever error. The static probability model gets almost all districts slightly wrong, only being accurate in the rare district that never changes during the entire model run. The simpler modal model reads as more accurate for the zones that don’t change during the out-of-sample period, but is glaringly wrong for districts that permanently transition away from the expected status. The carry forward model error is, again, associated with transitions and shows that a small number of front-line districts account for most of the districts. The ABM gets more wrong than the carry forward model, but has the lowest maximum error of any of the four models.

**Calibration and Discrimination**

Besides measuring the correctness of our forecasts directly using the Brier score, we can also decompose that measure into two different components: calibration and discrimination. Calibration is how close predicted probabilities are to the observed probabilities, and is defined as the difference between the forecasted likelihood of an event occurring and its real-world base rate. The left side of Figure 7 below shows the calibration of our forecasts. All model forecasts are binned into ten categories, and then within each bin we calculate the likelihood of those forecasts actually occurring. For example, out of all of our forecasts, we predicted an 80% to 90% likelihood of an event occurring 208 times and 197 of those events actually happened (94.7%). This is an overprediction and we can see that in the calibration graph that shows the 9th point above the “perfect calibration” line.

Discrimination on the other hand is calculated as the sum of the squared differences between the observed probabilities in each global forecast category (using the same bins as calibration) and the global base rate. A model could achieve high calibration by always predicting the global average, but since the observed frequency would be equal to the global frequency (as all samples are in the same bin), the model receives a poor discrimination score indicating its failure to give the user any meaningful information. Based on the right side of Figure 7 below, the ABM beats the modal and static models on both measures, but loses to the carry forward model on both measures.

---


Figure 7 - The left side graph shows the calibration of our model, with each point representing a bin where forecasts within each range are made. The x-axis shows the model predictions and the y-axis shows the true probabilities. The diagonal line represents where a perfectly calibrated model would fall. The right side graph compares our model output on both calibration and discrimination. Points in the upper-right are better, and the diagonal lines represent an even trade-off between calibration and discrimination.

Drilling Down

We may also wonder which particular districts are most inaccurate according to our ground truth data. As noted above, much of our forecasting error is concentrated in just a few districts. Figure 8 below shows the individual forecasts for all districts where our prediction of most likely control was wrong during any out-of-sample timestep.

In many districts our model predicts transitions that have not occurred, including that Iraq would take Khanaqin from the Kurds, that ISIS would conquer Tal Abyad, and that Assad would retake Al Mukharam. In others, our model predicts a continuation while the situation on the ground changes, such as Al Hasakah, Durna, and Muhardah. Muhardah is a good illustration of how this sort of error happens, where we miss a transition for Al Nusra, which has such a small footprint it practically never appears in the seeding or steering data. As a result, it’s common for runs to have no agents under the al Nusra zone, preventing the zone from having any role in the out-of-sample period. While the other zones are always present, the chaotic map of the Syrian conflict has many small isolated pockets of control that are difficult for agents to maintain and easily vanish.

Our model also predicts a transition late in the case of Izra and early in the case of Salamiyah. We would not expect our model to forecast the particular transition weeks, and an early warning may even count as a kind of success.
Figure 8 - Notable inaccurate forecasts. The lines are the forecast, and the background color is the ground-truth. Thicker lines indicate that the plurality forecast is incorrect.

Method 4 - Transition Rank Correlation

Although our model failed the t-test hypothesis that our model transition counts were the same as the world as it actually happened to unfold, we can still measure how different and in what way our transition likelihoods are incorrect. Using this method we rank each district by the number of transitions in the real-world, and the average likelihood of transition in our model. We then compare those rankings to see how closely our model ranking matches the world ranking. We can compare them statistically by using the Kendall Tau Rank Correlation\(^7\). We find a tau correlation of .38, and Figure 9 visualizes that correlation.

\[\text{Kendall's rank correlation tau} = 0.3826363\]

\[\text{alternative hypothesis: true tau is not equal to 0}\]

\[^7\text{https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient}\]
Figure 9 - Above are all of the districts ranked by likelihood to transition in the model (left) and the real-world (right). The lines connect each district, so straight lines indicate a correct ranking, and sloped lines indicate an incorrect ranking. This visual gives an approximation for the process the Kendall’s Tau metric, but does not take into account ties or paired rankings.

Cut Point Parameter Sensitivity Testing

In an additional set of experiments, we tested six different cut points for our in- and out-of-sample periods. We chose cut points at 30%, 40%, 50%, 60%, 70%, and 80% of transitions in the dataset (80% was the default used above.) Since our cut points were chosen as percentiles as the total number of transitions, they do not occur in equally-spaced intervals over time and are highly sensitive to the data. Figure 10 below illustrates those cut point percentages.
For our next set of experiments, we defined in-sample datasets ranging from 30% through 80% of the total transitions in increments of 10%. Since the transitions often spike, the horizontal lines could not be evenly spaced.

We ran 200 runs each for these five new cut points, and then measured the Brier scores over time for each experiment and compared their performance. Below in Figure 11 we can see that as expected, a later cut point improves performance of the model, although the shape of the curves tends to be consistent between experiments. The right side graph shows that if our forecast period starts at the 30% cut point (December 28th, 2014), our model still achieves a reasonable Brier score of about .03.

In order to gauge that Brier score, we also compared each new cut point model to the null hypothesis models within the same time period. Since the out-of-sample forecast period changes, our model forecasts and the null hypothesis baseline forecasts both change.
Figure 12 - The figures above compare our model’s Brier score during the final timestep (March 18th, 2016) to the three null hypothesis models under each of six scenarios.

Figure 12 above shows that while the carry forward null hypothesis model does beat the ABM’s mean Brier score when using the 80% cut point, the carry forward model begins to lose its advantage as the cut point is moved backwards in time, as seen in the 60% and 70% cutoff graphs. When the cut point is set to 30%, the carry forward model places last. This shows that the carry forward model performs relatively better in the short-term.

Figure 13 - Above we see the calibration-discrimination plot for the 30% and 50% cut point models (left and right, respectively.) We can see that the carry forward model loses both its calibration and discrimination between these cut point changes.

---

8 A t-test strongly suggests that the 30% cut point ABM and the carry forward model have different average Brier scores, with the p-value at 0.057, just short of the typical 95% confidence mark.
In Figure 13 above, we decompose the Brier score again into calibration and discrimination. We can see that moving the cut point from December 28th, 2014 (30%) to February 8th, 2015 (50%) considerably damages the carry forward forecasts. Interestingly, the static model actually performs best when making forecasts over a year into the future. The ABM results however continue to keep a steady calibration and discrimination. For the 60%, 70%, and 80% cut points the carry forward model outperforms the ABM, but the average ABM Brier score outperforms the null hypothesis models in the 30%, 40%, and 50% cut point scenarios (with varying levels of statistical significance). Further analysis into the comparison between each of these cut points and their corresponding null hypothesis models could be done in future work, and may illuminate in more detail where the tradeoffs lie when choosing between different model types.

Future Forecasts

While the in- and out-of-sample periods are useful for testing and analysis, the model was designed to make future forecasts. Below in Figure 14 are selected control charts for districts which our model predicts will have a transition between October 4th, 2016 and March 18th, 2018. Palmyra is an interesting case because while ISIS still controls the majority of territory in the district, Assad recently captured the city (and headlines) suggesting our model is on the right track.

![Territory Control Predictions](image)

**Figure 14 - Selected district futures showing predicted transitions within the next two years. The colored areas show to the ground-truth data up to the present.**
Ablation Study

Another way to examine the validity of the model is to disrupt key mechanisms and observe the resulting changes in the output. This can be used to test the robustness of forecasts, necessity of certain features, or the contribution of a mechanism to accuracy or error. We tested three ablations, removing from the model acquire candidates, elites, or zone rationality and examining the out-of-sample accuracy.

Ablate Acquire Candidates

In the first ablation study, we removed agents’ ability to change their subscription, making the model more primordialist in its treatment of identities. In Figure 15 we can see that the model performed about as well when the 80% cut point was used but did not outperform the static model when the 30% cut point was used (Brier score about .04). This is likely because removing the ability for agents to acquire new identities leads to a world where the status quo prevails and the model output more closely tracks the carry forward model. This also explains why when comparing the 80% cut point ablation ABM to the baseline ABM, the ablation model outperforms the average Brier score.

Figure 15 - Above we can see the average Brier score model comparisons for the 30% cut point (left) and the 80% cut point (right). The null hypothesis models are the same as in previous comparisons, but the ABM is now the acquire candidate ablation variation.

However in addition to the territory control forecasts, we can also compare the baseline and ablation models on other metrics such as DPH level size and average levels of attack (see Figure 16 below). Although we do not have real-world data to compare here, we can see that differences between the models, especially the average number of ISIS attacks, might cause us to wonder about the external validity of our model on other measures.
Ablate Elites

Without elites or elite networks, all agents in our model are identical, with an influence of one and a sight range of one. Networks are a tool of expansion by identities so we would expect this change to generally weaken Iraq and Syria State, Kurds, ISIS, and the Free Syrian Army. However, previous investigations have demonstrated that networks sometimes operate in reverse, as a path of infiltration by smaller identities, so we could not discard the possibility that removing networks may in some cases have the opposite effect.

We found that in both the 30% and 80% cut point cases, there were few discernible differences between the baseline and ablation experiment (on territory control predictions and other outcomes). From these results, we must conclude that either elites and networks have no effect on the model forecasts, or that there is something noteworthy about the Syraq context. The first can be discarded as initial development involved testing a great variety of different network configurations and examining the wide variety of outcomes for the most accurate representation. The alternative explanation is that in tuning the model to the chaotic, near-anarchic conflict of Syria, we have chanced upon a network and elite configuration that
broadly acts as little network at all. While some of the low-level dynamics have changed, the net effect of networks as support for identities and as infiltration-enabling weaknesses is close to zero. Because the network had such little influence on the model outcomes, we should continue to explore the way we operationalize elites by asking what level of influence elites have in the real-world and whether our model captures that reality.

**Ablate Zone Randomness**

Under this experiment, if a neighboring zone would place an agent in a higher political level, then it gets a stronger bonus for switching than under baseline circumstances. This weakens the effect of random probability and of incidental contexts such as recent attacks, and strengthens the rationality of individual agents. We found that the model does not outperform the baseline model, but does surprisingly well in the 30% cut point context (see Figure 16 left.) However it performs worse in the 80% cut point experiment (see Figure 16 right.) One way to interpret these results is that in the short-term this model takes too many risks by defining the state space of the possible too broadly, however that same risky strategy pays off in the long-term forecast situation. Figure 17 shows that in the 30% cut point experiment, even during steering the average zone size is less well determined because agents are constantly vying to be in the zone that presents the political opportunities with the highest payoff. Although we found that this model still did not beat our baseline forecast, further tuning of our “rationality” parameter might be fruitful in order to capture a more accurate political dynamic that favors agents’ internal preferences over the effect of external violence.

![Brier Scores Over Time, three null hypothesis models](image1)

![Brier Scores Over Time, three null hypothesis models](image2)

**Figure 16** - Above we can see the Brier score over time in our zone randomness ablation for a 30% cut point (left) and 80% cut point.
Figure 17 - Above we show the average zone size with 95% confidence intervals over time in our baseline and ablation experiments with a 30% cut point. This shows that the state space of the possible is larger in our ablation study, even during the steering period.

Conclusion

Tests conducted for this validation study of the ABM Syraq model sought to use one and only one data set for purposes of ground truth — the district control maps posted on Wikipedia. That meant, of course, that most of the output of the model was ignored: output pertaining, for example, to types of mobilization (violent or otherwise) by different groups, against different groups, at different places and different times, or pertaining to the rise and decline in the political fortunes of different groups within the core areas of their zones of predominance, or pertaining to the sequences of events leading to or preventing district control transitions. Metrics to evaluate the validity of the model with these outputs in mind would require different data and different approaches.

The tests conducted here reveal that the range of outcomes forecast by the model is not narrowly defined around the actual history of Syria and Iraq as measured by the Wikipedia maps, and that a small number of districts experienced transition patterns very different from those registered in our real-world data. On the other hand, the general trajectory that Syraq followed (featuring a relatively stable array of political competitors trading control over territory in the marches between them) was well within the central tendency of the model’s output. This is reflected in the Brier scores of the model when compared to the competing null hypothesis candidates and in the Kendall’s Tau assessment of the ranking of districts by their likelihood of experiencing transitions.

The ablation studies were extremely interesting and suggest the value of further work along these lines, not only to assist in model validation, but to help choose among competing ideas for operationalizing key variables.