Understanding convective mode as a predictor for warnings, issue hours, and lead times

Background and Research Question

There is a common misconception that tornadoes are not a threat outside of Tornado and Dixie Alley. For Tennessee residents, it is important to understand that all of Tennessee is at risk for tornadoes. More importantly, Tennessee residents are particularly vulnerable to tornadoes. This is a result of several demographic and social variables such as poverty and housing type. For this reason, meteorologists are key in providing up-to-date and understandable information regarding storm threats.

The goal of this project is to determine the relationship between convective mode and several common factors related to storm type, including (but limited to) false alarms, warnings, lead time, path length, time of day, and time of year. This data is specific to the Nashville county warning area (CWA). The warning area includes nearly forty counties in central Tennessee. This analysis intends to identify which properties of certain convective modes (cells in lines, cells in clusters, quasi-linear convective systems, and discrete supercells) affect the ability of forecasters to accurately predict tornadoes and issue warnings for the Nashville CWA.

Data

The data for this proposal come from Iowa State Mesonet and the SPC. All tornado path data were downloaded from the SPC and then compared to data from Iowa State Mesonet to determine lead times and warned/unwarned tornadoes. All false alarm data was downloaded using Iowa State Mesonet's API. Convective modes (storm type) were identified for all tornadoes and false alarms using GR2Analyst, a software used to read radar. - Radar data comes from NEXRAD on AWS: https://s3.amazonaws.com/noaa-nexrad-level2/index.html. - SPC: https://www.spc.noaa.gov/gis/svrgis/ - Iowa State Mesonet: https://mesonet.agron.iastate.edu/cow/

```
## ── Attaching packages ─────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1     ✔ purrr   0.3.3
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   1.0.2     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

Here is a preview of the raw data from "raw_torn_data." All datasets are similar to this format:

```
##     ID KellyNum UTCmonth UTCday UTCyear               valid season  CSTtime
## 1   1      343        1     17    2012 2012-01-17T19:27:00 winter 19:27:00
## 2   2      345        2     29    2012 2012-02-29T21:46:00 winter 21:46:00
## 3   3      346        2     29    2012 2012-02-29T22:02:00 winter 22:02:00
## 4   4      347        2     29    2012 2012-02-29T22:30:00 winter 22:30:00
## 5  18      354        3      2    2012 2012-03-02T21:48:00 spring 21:48:00
```

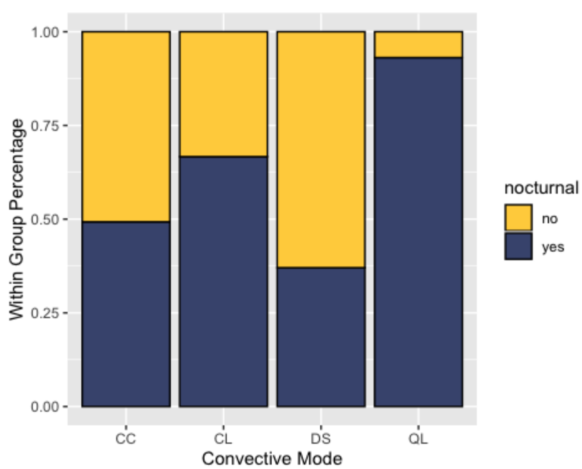
Methods

The overwhelming majority of the data are categorical variables. The focus of research is on the relationship between convective modes, warnings, false alarms, lead time and climatology (such as time of day and time of year). For this reason, models that adapt to the limitations of categorical values were used, such as logarithmic regressions and chi-squared tests to determine significance and model predictability. For numerical data, regular regressions were used.

Findings

Convective Mode and Nocturnality

The first statistical analysis is on convective mode and nocturnality. Is nocturnality dependent on convective mode or is the relationship random? For this analysis, the null hypothesis is the latter. First, the data were visually compared using a stacked bar chart. Then, a chi-squared test was performed.



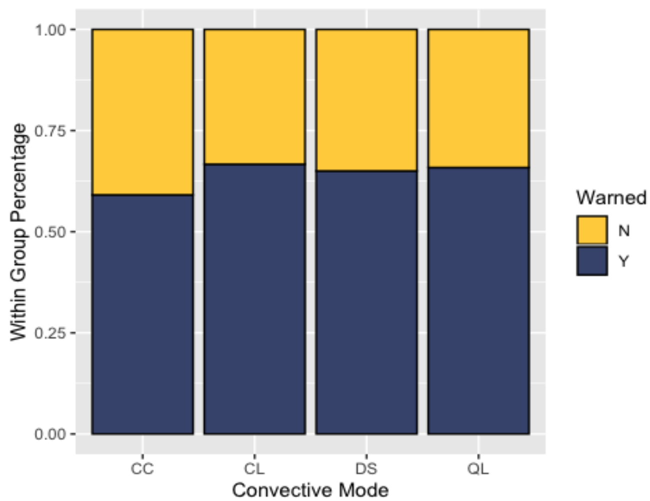
Note that the above graph was normalized to better understand the differences between convective modes. Visually, it can be assumed that some data associated with convective modes are not random, such as with quasi-linear convective systems (QL) and discrete supercells (DS). Most QL storms are nocturnal, while most DS storms occur during the day. This leads to my null hypothesis that there is not relationship between convective mode and time of day.

```
##
##  Pearson's Chi-squared test
##
## data:  raw_fa_data$mode and raw_fa_data$nocturnal
## X-squared = 22.595, df = 3, p-value = 4.905e-05
```

There is a significant relationship between mode and time of day accoring to the p-value, which is close to zero. The X-squared value is relatively high.

Convective Mode and Warnings

The next relationship, convective mode and warnings, was tested using a stacked bar chart and chi-squared test.



Visually, there appears to be no relationship between mode and warned. They are all very close. It is important to note that tornadoes were usually warned for in the Nashville area among all convective modes. Cell in cluster storms may be harder to warn for. To approach this, a chi-squared test was performed.

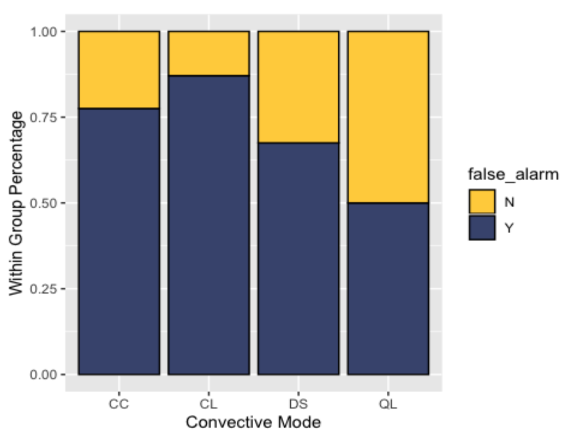
My mode data is nominal (4 groups), and my warned data is dichotomous (0, 1)

```
## Warning in chisq.test(raw_torn_data$Warned, raw_torn_data$mode): Chi-
## squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  raw_torn_data$Warned and raw_torn_data$mode
## X-squared = 0.32756, df = 3, p-value = 0.9548

##
## Call:
## glm(formula = Warned ~ mode, family = "binomial", data = raw_torn_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max  
## -1.4823  -1.4490   0.9140   0.9282   1.0258  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3677     0.4336   0.848    0.396
## modeCL        0.3254     0.8295   0.392    0.695
## modeDS        0.2513     0.6386   0.394    0.694
## modeQL        0.2891     0.5445   0.531    0.596
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 120.09  on 91  degrees of freedom
## Residual deviance: 119.77  on 88  degrees of freedom
## AIC: 127.77
## 
## Number of Fisher Scoring iterations: 4

## [1] 0.02221893
```

First, the p-value given from the chi-squared test is very high at .9548, so I cannot reject the null that the data is random. The logistic regression shows that there are no significant relationships between convective mode and whether or not storm was warned. Additionally, the p-value associated with the null deviance is .02, which means that convective mode is likely not the only variable that affects wheter or not a tornado was warned before touchdown. Overall, the relationship between convective modes and whether or not a tornado was warned is likely random. However, something other than convective mode may affect whether or not a tornado was warned.

Convective Mode and False Alarms



The graph above shows that quasi-linear storms produce the least amount of false alarms, while cell in line storms do. This shows that forecasters are likely to over-warn for cell in line, cell in cluster, and discrete supercell storms.

```
##
##  Pearson's Chi-squared test
##
## data:  all_warnings$mode and all_warnings$false_alarm
## X-squared = 17.783, df = 3, p-value = 0.0004876

##
## Call:
## glm(formula = false_alarm ~ mode, family = "binomial", data =
all_warnings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0237  -1.1774   0.7135   0.8866   1.1774
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2384     0.2540   4.876 1.08e-06 ***
## modeCL        0.6712     0.5929   1.132 0.257629
## modeDS       -0.5075     0.4224  -1.201 0.229622
## modeQL       -1.2384     0.3653  -3.390 0.000699 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 267.35  on 217  degrees of freedom
## Residual deviance: 249.53  on 214  degrees of freedom
## AIC: 257.53
##
## Number of Fisher Scoring iterations: 4
```

First, the results from the chi-squared test shows that there is some kind of relationship between convective mode and whether or not a storm produces a false alarm. This is because the p-value is at .0004876 which is well below .05. To further understand that, the logistic regression shows that there is a significant relationship between false alarms and convective modes for quasi-linear storms and cell in cluster storms. Specifically, by being a quasi-linear storm, this lowers the log odds of a false alarm by -1.24, which means that the odds are about 1/4th less than that of a cell in cluster storm. These results show that there is some significant relationship between convective mode and false alarms.

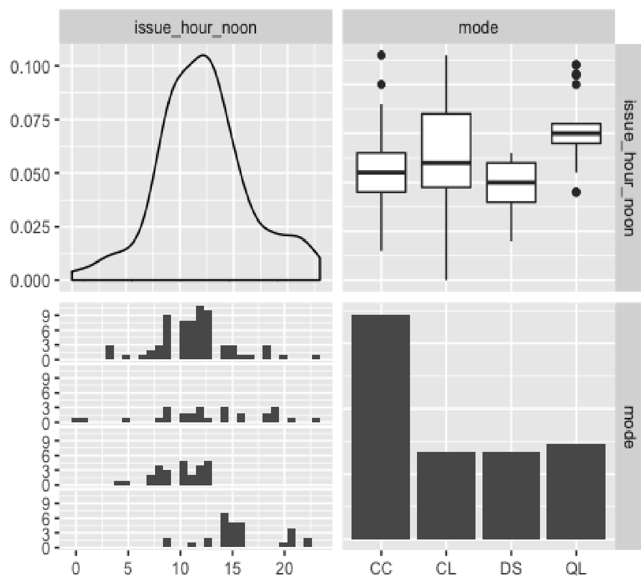
Convective Mode and Issue Hour

```
##
## Call:
## lm(formula = issue_hour_noon ~ mode, data = raw_fa_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7037  -1.8889  -0.3913   1.7343  11.6087
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3913     0.4688  24.298  < 2e-16 ***
## modeCL        1.3124     0.8840   1.485   0.1398
## modeDS       -1.5024     0.8840  -1.700   0.0913 .
## modeQL        4.3328     0.8618   5.027 1.42e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.894 on 148 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1824
## F-statistic: 12.23 on 3 and 148 DF,  p-value: 3.415e-07
```

What I get is that cell in line storm types are not significant, but quasilinear systems have a strong relationship with being nocturnal. Discrete supercells have a weak relationship with being during the day. Cell in Cluster storms graphically are 50/50, but the significance level is high here. This fits well with the assumption that cellular storms do what they want, discrete storms happen in the afternoon, and QLCS storms occur at night in the Southeast.

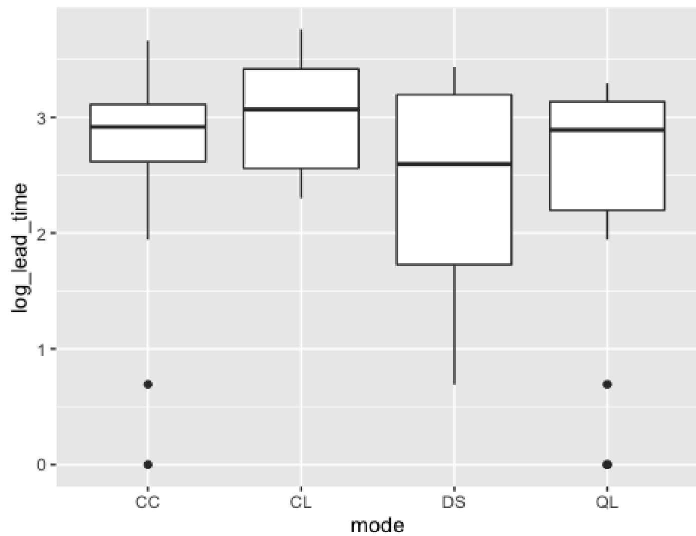
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



^ What did I do here? I had to adjust the time scale from showing 1 UTC -> 23 UTC to showing 12pm UTC -> 12am UTC. This is because otherwise it shows two peaks, when reality there is one overnight. This can affet the outcome of the data, so it is necesary to change this. The graph on the top left shows the peak in storms in the late afternoon to early morning. Interestingly, quasi-linear storms have a very tight boxplot showing that the usually happen at a certain time (night). The bottom right graph shows that most false alarms in the Nashville area are cell in cluster-type storms.

Convective Mode and Lead Time

```
## Warning: Removed 37 rows containing non-finite values (stat_boxplot).
```



First, for background, a log of 0 are tornadoes with 1 minute of lead time. A log of 1 is about 3 minutes. A log of 2 is about 9 minutes and so on. The above boxplot was useful in showing the wider range of lead times for discrete supercell tornadoes. It has the most variability. Cell in cluster storms on the other hand tend to have less variability. I took a log of the lead times because all the data was pushed down to the bottom without it (due to the number of lead times with zero minutes). Quasilinear storms must have a few storms that have very high lead time, considering that the median is pushed up so far even though the bulk of the data falls below the median. This might insinuate that quasilinear storms are harder to predict that the others. It is also important to note the outliers for clustered storms and quasi-linear storms, meaning that those types produce storms with very little lead times unlike cell in line storms and discrete supercells.

```
##
## Call:
## lm(formula = lead_time ~ mode, data = adj_torn_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5714  -7.6800   0.4286   7.1400  21.4286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.571      2.592   6.778 1.22e-08 ***
## modeCL         5.929      4.733   1.253    0.216
## modeDS        -2.571      4.016  -0.640    0.525
## modeQL        -2.211      3.238  -0.683    0.498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.7 on 51 degrees of freedom
## Multiple R-squared:  0.0701, Adjusted R-squared:  0.0154
## F-statistic: 1.282 on 3 and 51 DF,  p-value: 0.2907
```

The linear regression for the tornado data (minus the tornadoes not warned for – zero lead time) is very weakly positive. According to the p-values, there is insufficient evidence to reject the null hypothesis. It is possible that the relationship between lead times and convective modes is random, or at least not linear.

Limitations

One obvious limitation to the data is the time period. The data only represent 2012 to 2018, which is not a lengthy time period, especially when determining climatologies. Usually, a thirty year time period is best. This high number could not be achieved because convective modes have to be objectively and manually identified, which is extremely time consuming. Also, data post-2007 are very different from data pre-2007. Storm warnings used to be issued by counties, but now they are issued by polygons. Also, one must consider that not all tornadoes are spotted, especially EF0 tornadoes. Also, the identification of convective modes is subject to the person doing the identification. It is very difficult to perfectly separate each convective mode visually. Sometimes it is difficult to separate a cell in line from a quasi-linear system, or sometimes it is difficult to separate a discrete supercell from a cell in cluster system. Another pitfall is that the data is mostly categorical, which is more difficult to create predictions for.

Conclusion

In conclusion, I found that it is difficult to predict or relate certain variables by convective mode. Convective mode, in short, is not a very good explanatory variable. That is fair, mostly because identifying convective modes is very subjective, and the time period is very short. Also, there are so many other atmospheric variables that affect the response variables besides convective mode. That might be wind shear, wind direction, location, etc. What seems to be the most interesting convective mode is quasi-linear storms. In Nashville, they tend to be nocturnal, and they're difficult to predict. Cell in line storms in this area tend to produce the most false alarms in general, which is actually quite interesting.

References

1. Anderson-Frey, A.K., Y.P. Richardson, A.R. Dean, R.L. Thompson, and B.T. Smith. 2019. Characteristics of Tornado Events and Warnings in the Southeastern United States. Wea. Forecasting, **34**, 1017–1034. DOI: 10.1175/WAF-D-18-0211.1
2. Ashley, W.S. 2019. A Climatology of Quasi-Linear Convective Systems and Their Hazards in the United States. Wea. Forecasting, **34**, 1605–1631. DOI: 10.1175/WAF-D-19-0014.1

3. Brotzge, J.A., S.E. Nelson, R.L. Thompson, and B.T. Smith. 2013. Tornado Probability of Detection and Lead Time as a Function of Convective Mode and Environmental Parameters. Wea. Forecasting, **28**, 1261–1276. DOI: 10.1175/WAF-D-12-00119.1

4. Davis, J.M., and M.D. Parker. 2014. Radar Climatology of Tornadic and Nontornadic Vortices in High-Shear Low-CAPE Environment in the Mid-Atlantic and Southeastern United States. Wea. Forecasting, **29**, 828–853. DOI: 10.1175/WAF-D-13-00127.1

5. Ellis, K.N., D. Burrow, K.N. Gassert, L. Reyes-Mason, and M.S.Porter. 2019. Forecaster Perceptions and Climatological Analysis of the Influence of Convective Mode on Tornado Climatology and Warning Success. Annals of the American Association of Geographers, **109**, 1–20. DOI: 10.1080/24694452.2019.1670042