Using weather to predict West Nile Virus risk in Nebraska

BY KELLY HELM SMITH
WITH MUCH INPUT FROM ANDREW J. TYRE, JEFF HAMIK, MICHAEL J. HAYES, YUZHEN ZHOU, LI DAI
PUBLIC HEALTH ASSOCIATION OF NEBRASKA,
LINCOLN, APRIL 3, 2019

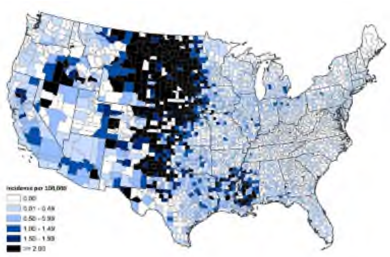
NDMC NATIONAL DROUGHT MITIGATION CENTER
UNIVERSITY OF NEBRASKA

1

Our question:

Can we use weather data to predict years with greater risk of West Nile Virus infection?

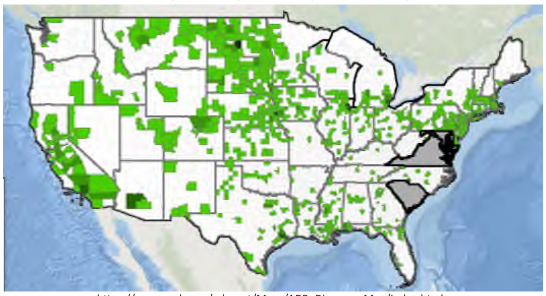
Average annual incidence of WNV reported to CDC by county, 1999-2017



https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

2

Human cases of WNV by county, 2018, preliminary CDC Arbonet data



https://wwwn.cdc.gov/arbonet/Maps/ADB_Diseases_Map/index.html

NATIONAL DROUGHT MITIGATION CENTER

3

Culex tarsalis

Culex tarsalis is the mosquito species most responsible for transmitting West Nile virus in Nebraska.

Infection season in Nebraska is roughly June – September.

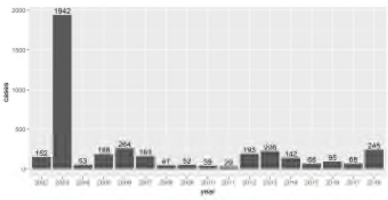
State monitors mosquitos in 30 counties. But we are starting analysis with human cases, due to greater spatial coverage.

Anecdotal observations: Drought contributes to higher rates of West Nile Virus in humans.

University of Nebraska photo

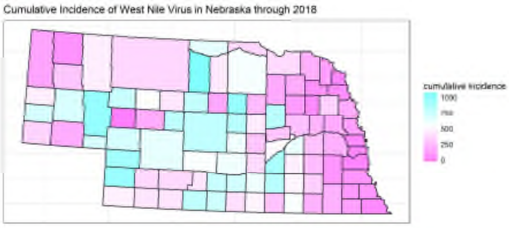
4

Human cases of West Nile Virus in Nebraska, 2002-2018



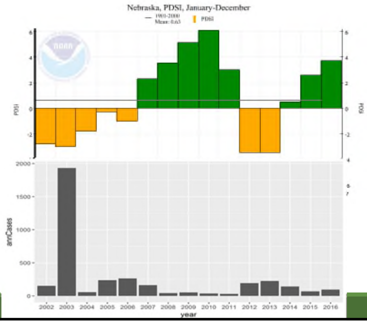
5

Cumulative WNV Incidence/100K through 2018



6

Drought and West Nile Virus by year in Nebraska



7

The equation (model) in words:

What we want to
1) explain and
2) predict, the
"Dependent variable" or
"Response variable"

The number of human cases of West Nile Virus in each year and county

"cases," for short

Unit of analysis: county-year

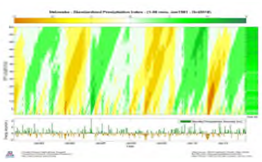
What we are using to make predictions, "predictors" or "independent variables"

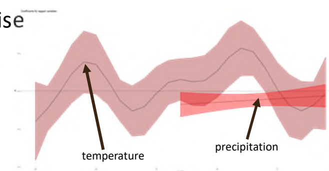
- Temperature
- A drought index: Standardized Precipitation (and Evapotranspiration) Index, SP(E)I
- Cumulative incidence (how many people have already had WNV)
- Population
- County (to account for unique spatial factors)
- Year (to account for unique temporal factors)

NATIONAL DROUGHT MITIGATION CENTER

8

Finding the signal in the noise





We looked at lags 12, 18, 24, 30 & 36-month lags for:
- temperature
- precipitation (departure from normal)
- Standardized Precipitation Index
- Standardized Precipitation & Evapotranspiration Index

Coefficients for the lagged temperature and drought variables indicated that warmer winters and drier years increased the number of West Nile Virus cases.

We made models based on data through 2017, and then used the best-performing model to make predictions for 2018 … And on data through 2018, to predict for 2019.

NATIONAL DROUGHT MITIGATION CENTER

9

A quick aside about statistics, AKA expectation management …

"All models are wrong, but some models are useful." – George Box

https://commons.wikimedia.org/wiki/File:GeorgeEPBox.jpg

Hirotugu Akaike, "Akaike's Information Criterion" (AIC), 1974

https://commons.wikimedia.org/wiki/File:Akaike.jpg

Practically speaking, AIC is a quick, efficient means of determining which of many models is most useful. In combination with modern computing capabilities and open-source software, such as R, this puts some fairly powerful tools at our disposal.

NATIONAL DROUGHT MITIGATION CENTER

10

Evaluating models

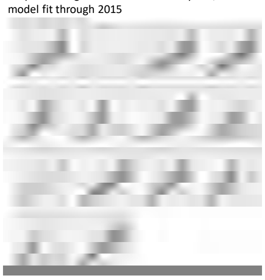
We used AIC to identify the best models among the 384 possible combinations of variables that we tested. We evaluated models that were within 2-4 points of the lowest AIC score, based on several criteria:

- Model criteria
  - R-squared
  - deviance explained
  - lack of spatial and temporal autocorrelation in the residuals
- Performance criteria
  - Ratio of predicted to actual cases, year-by-year
  - Ratio of predicted to actual cases, county-by-county
- Comparison with the naïve model, the assumption that a county will have the same number of human cases that it did the previous year. Does our model tell us something we don't already know?
  - By year
  - By county

NATIONAL DROUGHT MITIGATION CENTER

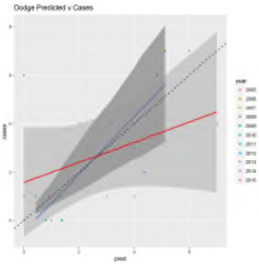
11

Calibration curve showing fitted model outperforming naïve in 10 out of 13 years, for model fit through 2015



We looked for fitted models with a ratio of predicted to actual cases that was near 1:1.

Dodge Predicted v Cases



NATIONAL DROUGHT MITIGATION CENTER

12

Calibrating understanding (getting a feel for how it works)

2002-16 + 2017 June	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves			unties in .5-1.5 target r		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	notes
mods_coyr2.rmd							2003-2016	2017		2005-2016	2017	2005-2016	2017	2003-2016		2017			
M1_t12spei24	3581.115787	3908.351	90.2	81	78.4	71.1	12/14, 86%	0/1		53%		81%	65%	25.6% v 9.8%	90.371, 1, < 2.2e-16	17.4% v 8.7%	2.0417, 1, 0.153	s(lags_tmean12, by = tmean12) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
M2_spei24	3582.654007	4101.888	87.9	60.2	78.1	62.9	9/14, 64%	1 out of 1		46%		70%	68%	24.8% v 9.2%	91.116, 1, < 2.2e-16	26.1% v 8.7%	7.0312, 1, 0.00801	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
M3_t24spei24	3583.433175	3895.446	88.9	84.3	78.2	71.9	10/14, 71.4%	0/1		58%		82%	62%	25% v 8.9%	97.108, 1, < 2.2e-16	18.5% v 6.5%	4.3478, 1, 0.03706	s(lags_tmean24, by = tmean24) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop1(	tmean not sig in refit
M4_t12spi24	3584.153641	3914.39	90.3	79.1	78.3	70.8	12/14, 86%	0/1*		50%		81%	65%	25.6% v 9.3%	96.356, 1, < 2.2e-16	16.3% v 7.6%	2.2273, 1, 0.1356	s(lags_tmean12, by = tmean12) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
M5_t30spei24	3585.092408	3881.768	89.2	86	78.2	72.60	11/14, 78.6%	0/1*		55.4%		82.2%	65.8%	25% v 9.2%	92.526, 1, < 2.2e-16	19.6% v 8.7%	3.1154, 1, 0.07756	s(lags_tmean30, by = tmean30) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop10(	p value of temperatu
2002-16 + 2017 May	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	notes
							2003-2016	2017	2005-2016	2005-2016	2017	2005-2016	2017	2003-2016		2017			
May_1	3588.565066	4075.23	88.7	65.2	78.1	64.2	7/14, 50%	1 out of 1	25%	49%		71.2%	72.2%	24.5% v 10.4%	72.802, 1, < 2.2e-16	23.9% v 5.4%	9.4815, 1, 0.002076	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
May_2	3589.741957	3892.481	90.2	79.9	78.3	71.9	11/14, 77%	0/1*	67%	57%		81.7%	63.0%	24.8% v 8.3%	104.51, 1, < 2.2e-16	17.4% v 10.7%	0.96154, 1, 0.3268	s(lags_tmean18, by = tmean18) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
May_3	3591.046938	3881.436	90.2	83.3	78.3	72.4	11/14, 77%	0/1	67%	59%		81.8%	62.5%	25.1% v 8.5%	103.8, 1, < 2.2e-16	18.5% v 9.8%	1.8846, 1, 0.1698	s(lags_tmean18, by = tmean18) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
May_4	3592.883407	3915	90.4	83.3	78.4	71.2	12/14, 86%	0/1*	83%	54%		82.4%	66.8%	24.5% v 8.1%	106, 1, < 2.2e-16	19.6% v 9.8%	2.3704, 1, 0.1237	s(lags_tmean18, by = tmean18) + s(lags_spi30, by = spi30) + CI + County + year + offset(log(pop100K))	
2002-16 + 2017 Apr	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance (original computation), Chi-sq, df & p				formula	
							2003-2016	2017	2005-2016	2005-2016	2017	2005-2016	2017	2003-2016		2017			
Apr_1	3577.87732	3878.142	90.8	75.5	78.8	72	9/14, 64%	*	58.3	46.7		80.9	75.2	24.8% v 8.3%	104.51, 1, < 2.2e-16	17.4% v 6.5%	3.6818, 1, 0.05501	s(lags_tmean12, by = tmean12) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
Apr_2	3578.199999	3867.863	90.7	76.8	78.8	72.3	11/14, 77%	*	58.3	48.9		80.9	74	24.7% v 8.5%	100.11, 1, < 2.2e-16	18.5% v 6.5%	4.3478, 1, 0.03706	s(lags_tmean12, by = tmean12) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
2002-16 + 2017 Mar	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	
							2003-2016	2017	2005-2016	2005-2016	2017	2005-2016	2017	2003-2016		2017			
Mar_1	3582.963584	3873.523	90.5	84	78.6	72.6	10/14, 71%	0/1	58.3	55.4		83.3	65.2	24.3% v 8%	105, 1, < 2.2e-16	18.5% v 9.8%	1.8846, 1, 0.1698	s(lags_tmean18, by = tmean18) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
Mar_2	3586.603614	3903.895	90.7	68.6	78.4	70.9	12/14, 86%	*	50	57.6		80	70.1	25.3% v 8.9%	100, 1, < 2.2e-16	17.4% v 6.5%	3.6818, 1, 0.05501	s(lags_tmean12, by = tmean12) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
Mar_3	3586.636218	4059.145	89.6	61	78.2	64.7	9/14, 64%	1/1*	25	46.7		66	65	25.7% v 9.9%	90.667, 1, < 2.2e-16	19.6% v 7.6%	4, 1, 0.0455	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
2002-16 + 2017 Feb	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	
							2003-2016	2017	2005-2016	2005-2016	2017	2005-2016	2017	2003-2016		2017			
Feb_1	3586.539039	4027.244	87.9	70.6	78.1	66.3	7/14, 50%	1 of 1	16.7	48.9		75.2	74.2	23.8% v 10.6%	65.237, 1, 6.641e-16	22.8% v 7.6%	6.0357, 1, 0.01402	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
Feb_2	3589.890373	3845.124	88.8	87.5	78.1	73.7	11/14, 77%	0/1	50	57		82.1	65.3	25.3% v 9.1%	96.943, 1, < 2.2e-16	17.4% v 5.4%	4.7619, 1, 0.0291	s(lags_tmean36, by = tmean36) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop1(	tmean not significant
Feb_3	3590.142373	3842.394	88.8	86.2	78.1	74	12/14, 86%	0/1	0.75	57.6		83.4	70.9	24.6% v 9.3%	87.908, 1, < 2.2e-16	16.3% v 7.6%	2.2273, 1, 0.1356	s(lags_tmean30, by = tmean30) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
Feb_4	3590.468	4096.199	88	66.2	78	63.2	6/14, 43%	1 of 1	33	47.8		72	72.1	24.4% v 10.6%	69.62, 1, < 2.2e-16	21.7% v 7.6%	5.3333, 1, 0.02092	s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
2002-17 + 18 Jun	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	
							2003-2017	2018	2005-2017	2005-2017	2018	2005-2017	2018	2003-2017		2018			
M1718_Jun_1	3749.374276	4224.563	89.7	75.3	77.8	68.1	12/15, 80%	1 of 1	46.2	51.1		79.4	94.1	25.4% v 10.3%	87.091, 1, < 2.2e-16	25% v 7.6%	7.5, 1, 0.00617	s(lags_tmean12, by = tmean12) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
M1718_Jun_2	3749.555272	4469.706	87.5	46	77.6	57.6	8/15, 53%	0/1*	38.5	42.4		68.8	90.4	24.1% v 11%	66.804, 1, 2.999e-16	20.7% v 7.6%	4.6538, 1, 0.03098	s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
M1718_Jun_3	3749.714911	4269.539	89.3	76.6	77.7	66.1	9/15, 60%	*	30.8	46.7		75.3	91.9	24.1% v 9.9%	80.904, 1, < 2.2e-16	22.8% v 9.8%	4.0333, 1, 0.04461	s(lags_tmean18, by = tmean18) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
M1718_Jun_4	3749.966355	4157.127	90.3	82.1	78	70.5	9/15, 60%	1 of 1	23.1	57.6		79.7	91.1	24.5% v 9.9%	84.211, 1, < 2.2e-16	29.3% v 10.9%	6.9189, 1, 0.008529	s(lags_tmean30, by = tmean30) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
2002-17 + 18 Feb	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	
							2003-2017	2018	2005-2017	2005-2017	2018	2005-2017	2018	2003-2017		2018			
M1718_Feb_1	3749.427966	4377.177	88.2	66.2	77.7	61.5	7/15,	1 of 1	38.5	43.5		71.7	91.9	24.1% v 11.4%	61.662, 1, 4.078e-15	23.9% v 7.6%	6.7586, 1, 0.00933	s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
M1718_Feb_2	3751.736461	4367.117	87.9	68.4	77.6	61.9	8/15,	1 of 1*	30.8	44.6		71.4	80.6	24.3% v 11.2%	65.988, 1, 4.537e-16	18.5% v 6.5%	4.3478, 1, 0.03706	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
M1718_Feb_3	3752.868164		88.7	81.9	77.7	70	10/15, 66%	1 of 1	53.8	51.1		78.2	91.1	23.8% v 9.1%	90.075, 1, < 2.2e-16	23.9% v. 10.9%	3.7812, 1, 0.05183	s(lags_tmean24, by = tmean24) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))	
2002-18 + 19 Feb	AIC	AIC refit	R^2	R^2 refit	deviance expl	dev refit	Calibration curves		% years in	% counties in .5-1.5 ta		Cor cases ~ pred		Performance, Chi-sq, df & p				formula	
							2003-2018	2019	2005-2018	2005-2018	2019	2005-2018	2019	2003-2018		2019			
M1819_Feb	4004.552344	4191.79	90.4	85.5	76.9	71.6	9/16, 56%		28.6	58.7		77.3		24.3% v 9.2%	98.868, 1, < 2.2e-16			s(lags_tmean36, by = tmean36) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))	
*very close																			

NATIONAL DROUGHT MITIGATION CENTER

13

Things we learned: New data changes the "answer"

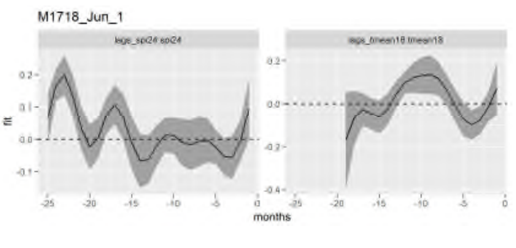
There is no single "right" model, but most of the top models used 24-month drought indicators, either SPI or SPEI. The best-fit model varies based on the time interval of the data that we're looking at. Note that models based on data through October used 30-month SPEI lags – going back further in time.

Model name	AIC	formula
M1718_Feb	3744.993163	s(lags_tmean12, by = tmean12) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))
M1718_Jun_1	3751.229582	s(lags_tmean18, by = tmean18) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))
	3751.259335	s(lags_tmean30, by = tmean30) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))
	3752.709596	s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))
	3753.118337	s(lags_tmean30, by = tmean30) + s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))
	3753.686899	s(lags_tmean18, by = tmean18) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))
	3753.979161	s(lags_spi24, by = spi24) + CI + County + year + offset(log(pop100K))
	3754.04474	s(lags_tmean12, by = tmean12) + s(lags_spei24, by = spei24) + CI + County + year + offset(log(pop100K))
M18_Oct_1	4042.174726	s(lags_tmean12, by = tmean12) + s(lags_spei30, by = spei30) + CI + County + year + offset(log(pop100K))
M18_Oct_2	4043.645099	s(lags_tmean36, by = tmean36) + s(lags_spei30, by = spei30) + CI + County + year + offset(log(pop100K))

NATIONAL DROUGHT MITIGATION CENTER

14

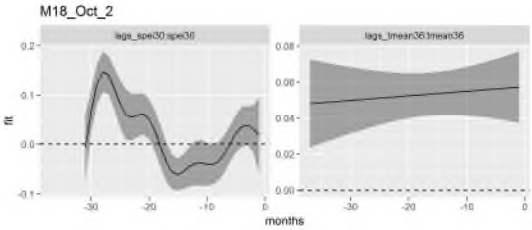
Long-term influence of drought & temperature, data through June 2018



NATIONAL DROUGHT MITIGATION CENTER

15

Long-term influence of drought & temperature, data through October 2018



16

Things we learned: AIC doesn't tell the whole story

Sometimes there was a lot of variation even between the top two models in a set. Two models were best, by AIC, through October 2018, but one was better at space and the other was better at time. Choosing the "right" model may depend on what you want to know.

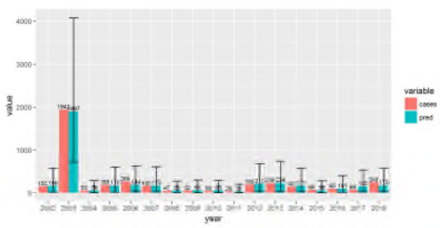
	AIC	Calibration curves	% years in .75-1.25 target range	% counties in .5-1.5 target range	Correlation cases ~ pred	Performance, Chi-sq, df & p		formula
		2003-2017	2005-2018	2005-2018	2005-2018	2003-2018		
M18_Oct_1	4042.175	12/16, 73%	57.1	45.7	70.1	25.1% v 10.5%	87.23, 1, < 2.2e-16	s(lags_tmean12, by = tmean12) + s(lags_spei30, by = spei30) + CI + County + year + offset(log(pop100K))
M18_Oct_2	4043.645	8/16, 50%	28.6	59.8	79.5	24.9% v 9.8%	95.767, 1, < 2.2e-16	s(lags_tmean36, by = tmean36) + s(lags_spei30, by = spei30) + CI + County + year + offset(log(pop100K))

17

Things we learned: Some years more predictable than others

Predicting numbers of cases is harder than predicting which counties will have cases, and some years are "easier" to predict than others, which may mean influences other than the weather were important.

Predictions vs. cases, based on M18_Oct_2 (data through October 2018, second model)



18

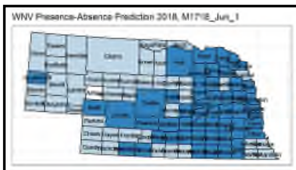
Things we learned: “Where” was easier than “how many”

Our weather-based models were *always* better at annual presence-absence predictions by county than the naïve model. This is particularly relevant, given that slightly more than half of Nebraska county-years 2002-2018 (828 out of 1564, or 53%) had zero human cases.

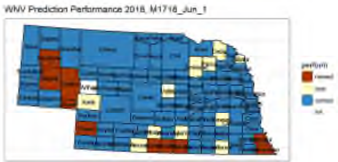
2002-17 + 18 Feb	Performance, Chi-sq, df & p			
	2003-2017		2018	
M1718_Feb	25.0% v 10.2%	84.792, 1, < 2.2e-16	23.9% v 10.9%	3.7812, 1, 0.05183
2002-17 + 18 Jun	Performance, Chi-sq, df & p			
	2003-2017		2018	
M1718_Jun_1	24.2% v 9.9%	81.563, 1, < 2.2e-16	23.9% v 10.9%	3.7812, 1, 0.05183
2002-17 + 18 Oct	Performance, Chi-sq, df & p			
	2003-2018			
M18_Oct_1	25.1% v 10.5%	87.23, 1, < 2.2e-16		
M18_Oct_2	24.9% v 9.8%	95.767, 1, < 2.2e-16		

A model based on data through June 2018 did better than the naïve model 24.2% of the time, whereas the naïve model did better 9.9% of the time, for 2003-2017, a difference that was very unlikely to happen by chance, according to McNemar’s statistical test of paired binomial data. For 2018 – the out-of-sample year – the model outperformed the naïve 23.9% of the time, and the naïve outperformed the fitted model 10.9% of the time, a difference likely to happen by chance just under 95% of the time.

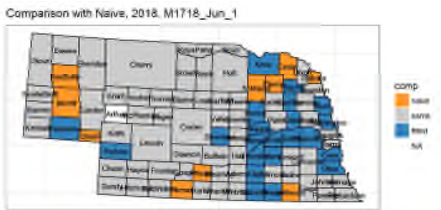
19



Presence-absence predictions & performance for 2018, based on data through June 2018



20



Our best-performing model based on data through June 2018 correctly predicted 23.9% of the counties in Nebraska in 2018 that the naïve model missed. The naïve model correctly predicted 10.9% of the counties that our fitted model missed.

21

Things we learned: It's worth a shot

In our explorations of 2017 and 2018, we found that predictions for the coming year based on data through February of that year can be nearly as accurate as predictions made in June, and even comparable to retroactive modeling, looking backwards from October.

This suggests that:

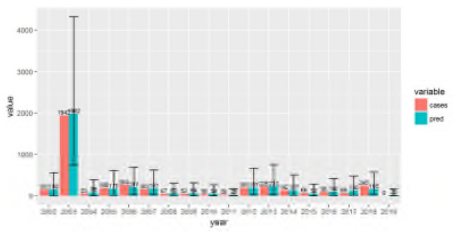
- The weather-climate contribution to human cases of WNV in NE is long-term – dry years preceded by wet years, with warm winters.
- This process can provide advance notice of years with increased risk of WNV.

… so with no further ado …

22

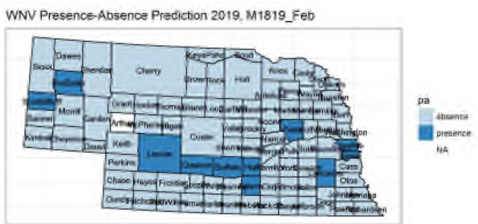
2019 Prediction

Applying this process to data through February 2019, we anticipate between 0 and 187 human cases of WNV. The literal prediction is 23, which would be the least number of cases since tracking began, but the 95% prediction interval provides a much larger margin, up to 210.



23

2019 Prediction

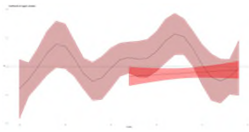


Cases in:

- Box Butte
- Scotts Bluff
- Lincoln
- Dawson
- Buffalo
- Hall
- Adams
- Platte
- Lancaster
- Douglas
- Sarpy

24

Questions, comments?



Please write or call:

Kelly Helm Smith

ksmith2@unl.edu

402-472-3373

NATIONAL DROUGHT MITIGATION CENTER

25