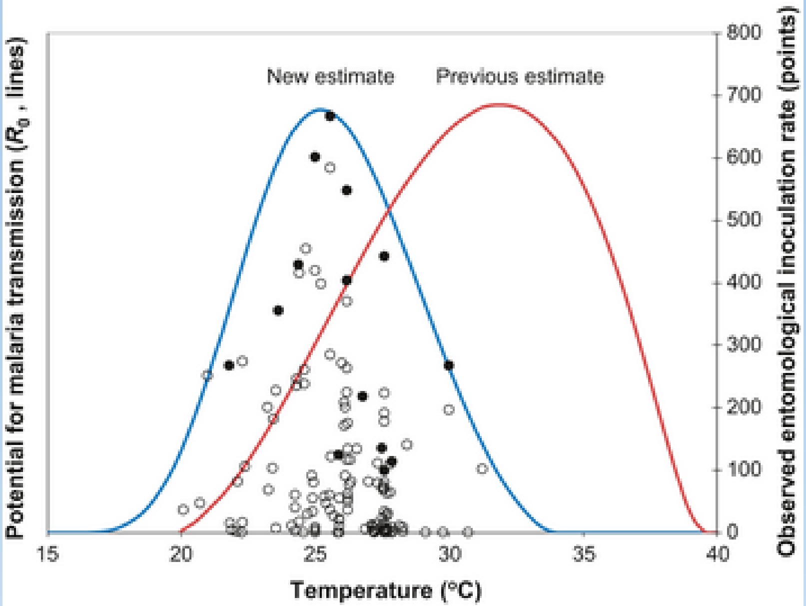
Spatiotemporal Bayesian Networks for Prediction of Vector-Borne Disease

Prof. Peter Haddawy
Faculty of ICT
Mahidol University

Challenges

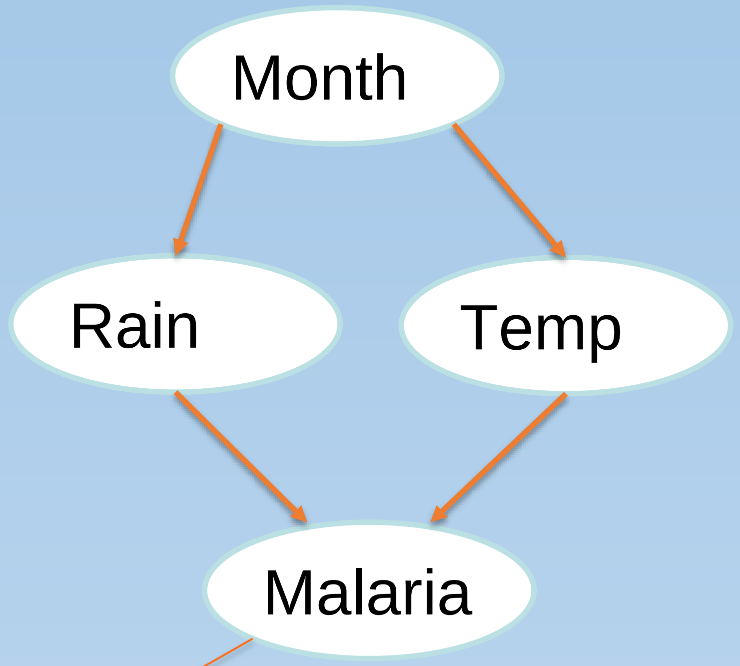
- Disease transmission can exhibit spatial and temporal heterogeneity, spatial autocorrelation, and seasonal variation
- Environmental factors can affect vector population and disease transmission in complex and non-linear ways
  - Temperature – vector & parasite maturation,
    biting rate
  - Rainfall – flushing effect
- Representing uncertainty



Mordecai et al, Ecology Letters 16:22-30, 2013

Bayesian Networks

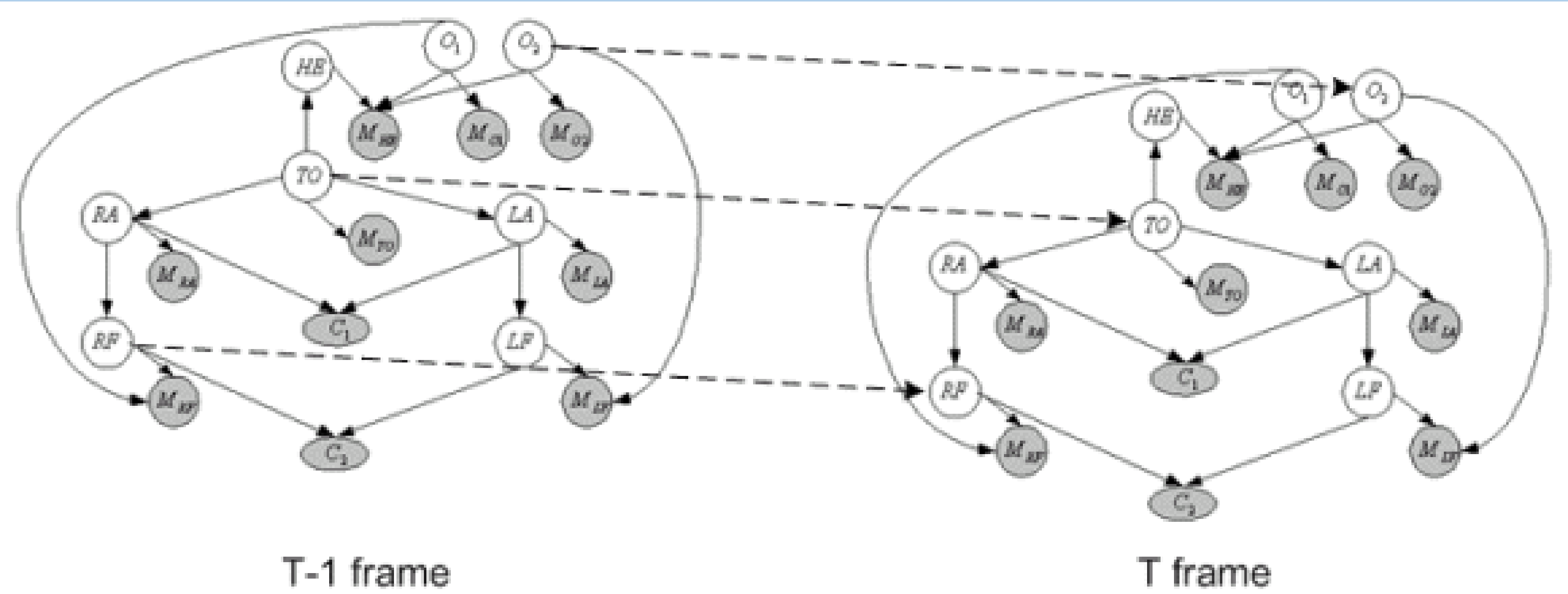
- Graphical representation of a probability distribution
- Nodes - random variables
- Edges – direct influence, quantified by conditional probability
- Probabilistic semantics: Malaria independent of Month given Rain and Temp



P(Malaria = H | Rain = H, Temp = H)
P(Malaria = H | Rain = H, Temp = L)
…

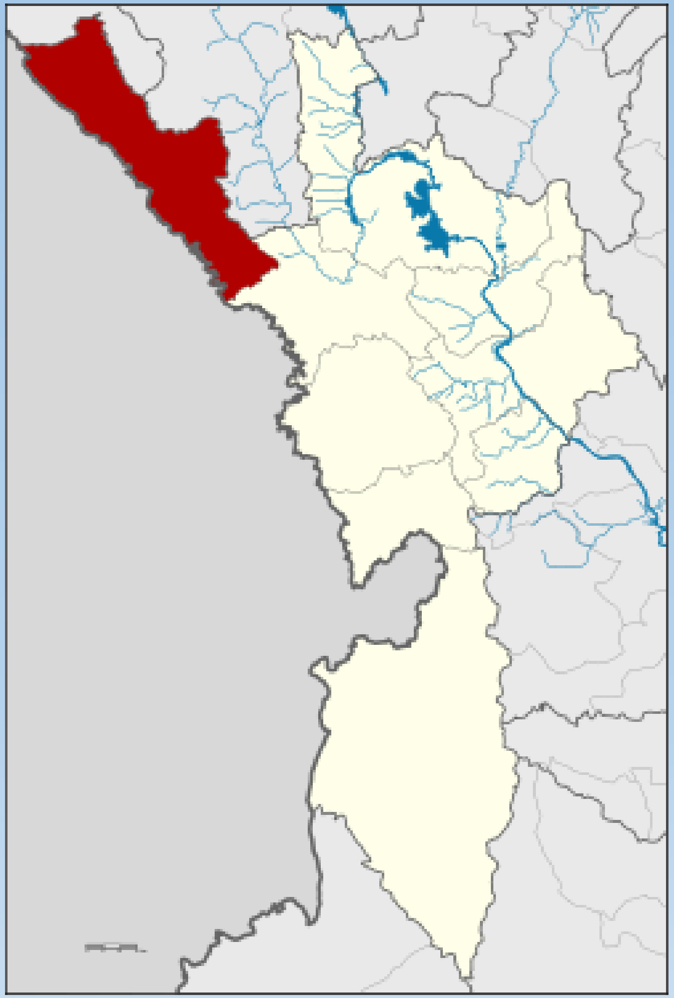
Dynamic Bayesian Networks

- A way to represent Markov models: HMM’s, Kalman filters, others
- Nodes organized into time slices
- States: Collections of nodes in a time slice and relations among them
- Transitions: Links between time slices



Tha Song Yang (ท่าสองยาง)

- 1,920 km^2 district in Tak province
- 66 villages near border with Myanmar
- Malaria is endemic
- Imported cases from Myanmar



Malaria Case Data

- Two years (2012 – 2013) of weekly clinically confirmed malaria case data from 66 villages
- 6,579 records with 12,800 total cases (PF, PV)
  - Cases: Min 0, Max 82, Mean 2.1

Environmental Data

temporal

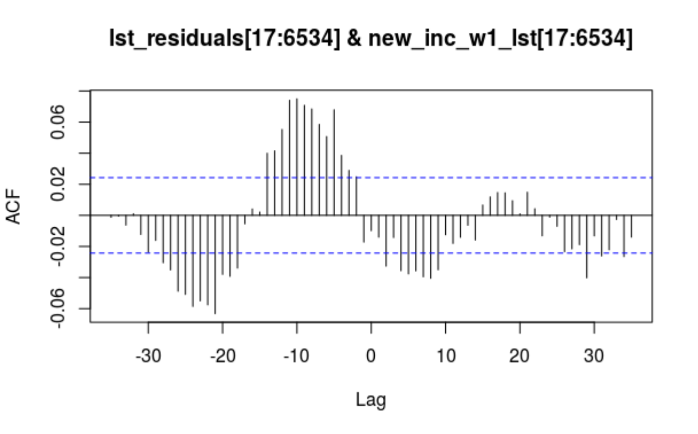
- NDVI
  - Monthly from MODIS (MOD11A3)
  - Temporal interpolation used to fill in missing values
- Rainfall
  - Daily at 10 km resolution
  - From JAXA Global Rainfall Watch
- Land Surface Temperature (LST)
  - Monthly at 5 km resolution from MODIS (MOD11C3)
  - Spatial and temporal interpolation to fill in missing values

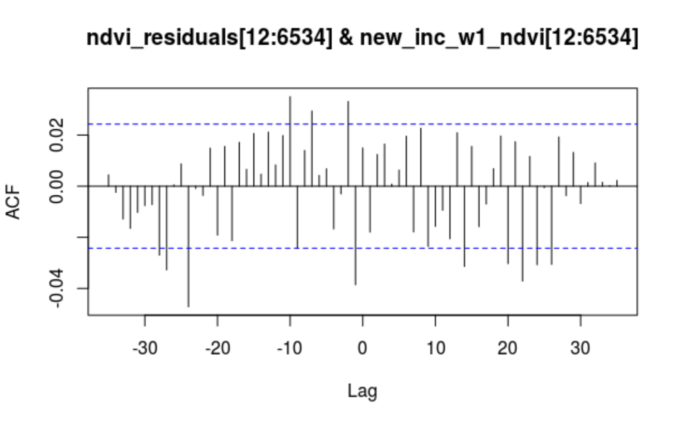
non temporal

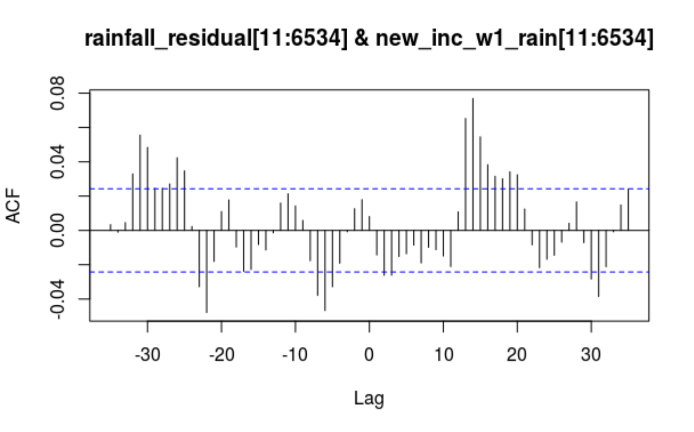
- Streams
  - Distance to closest stream
  - Stream density: Total stream length in 4 km buffer
- Slope
  - Average in 1 km buffer (from elevation data)
- Distance to border: proxy for imported cases

Model Building: Time Lags

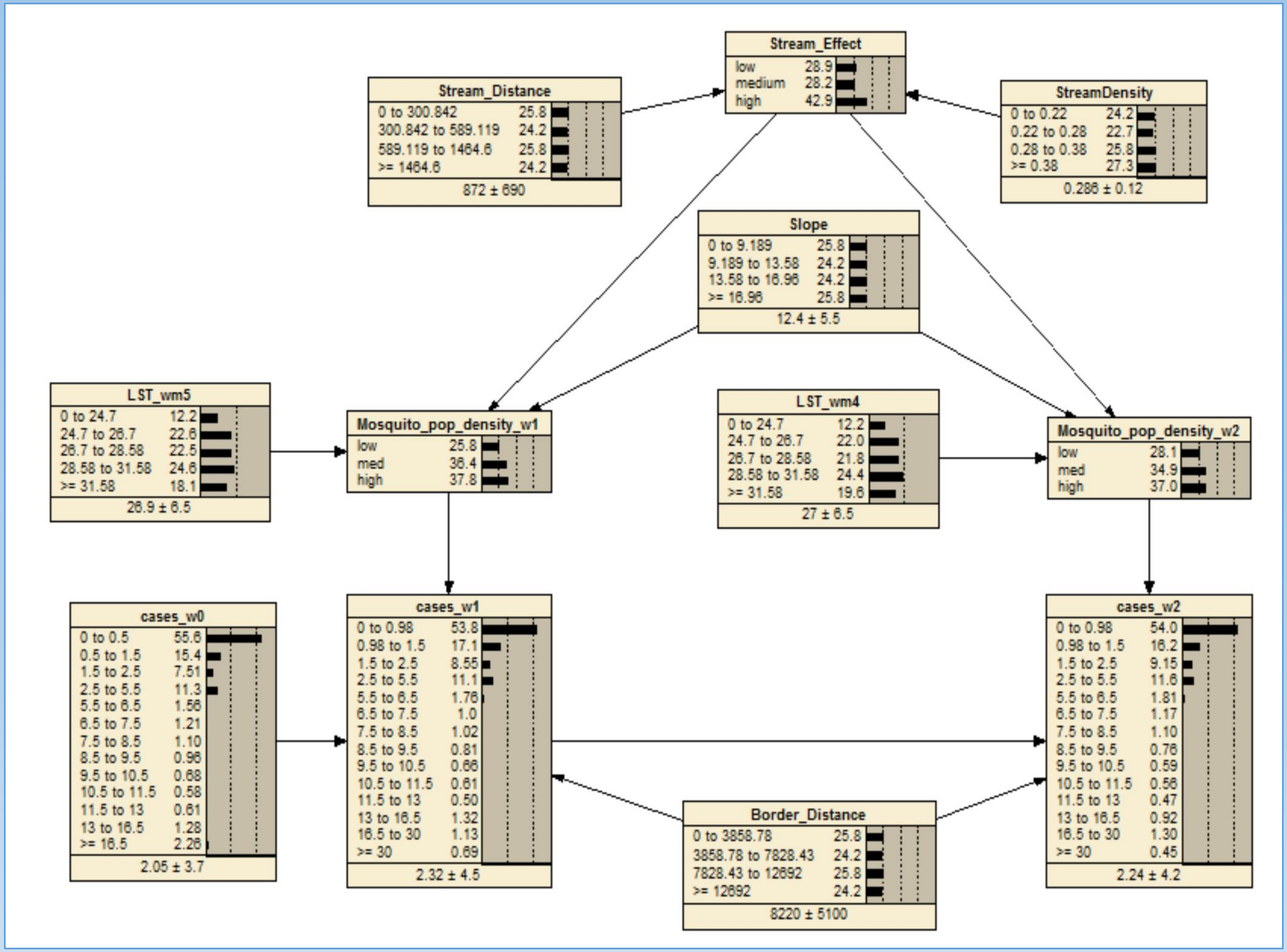
- Environmental variables effect different parts of the vector and parasite cycle
- Time lags determined using cross-correlation with pre-whitening + model fitting
  - Fit ARIMA model to independent variable X
  - Use to filter dependent variable Y
  - Calculate cross-correlation on residuals for X and filtered Y
- LST: 6 weeks
- Rainfall: 7 weeks
- NDVI: 8 weeks





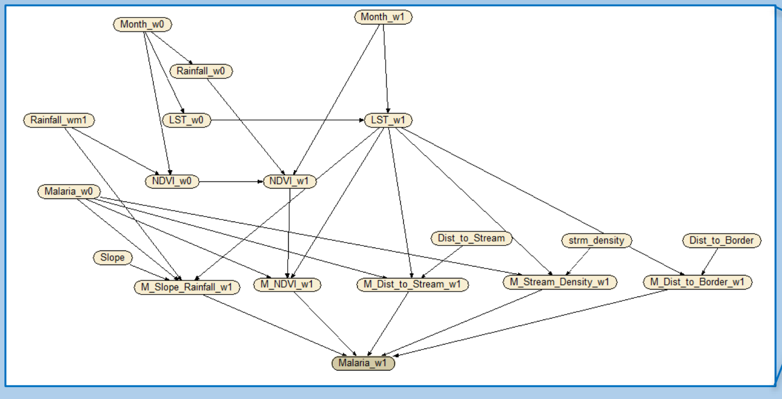


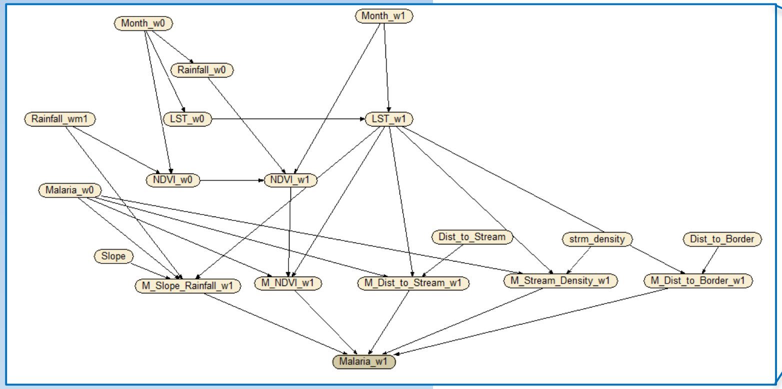
Village Level Bayes Net Model

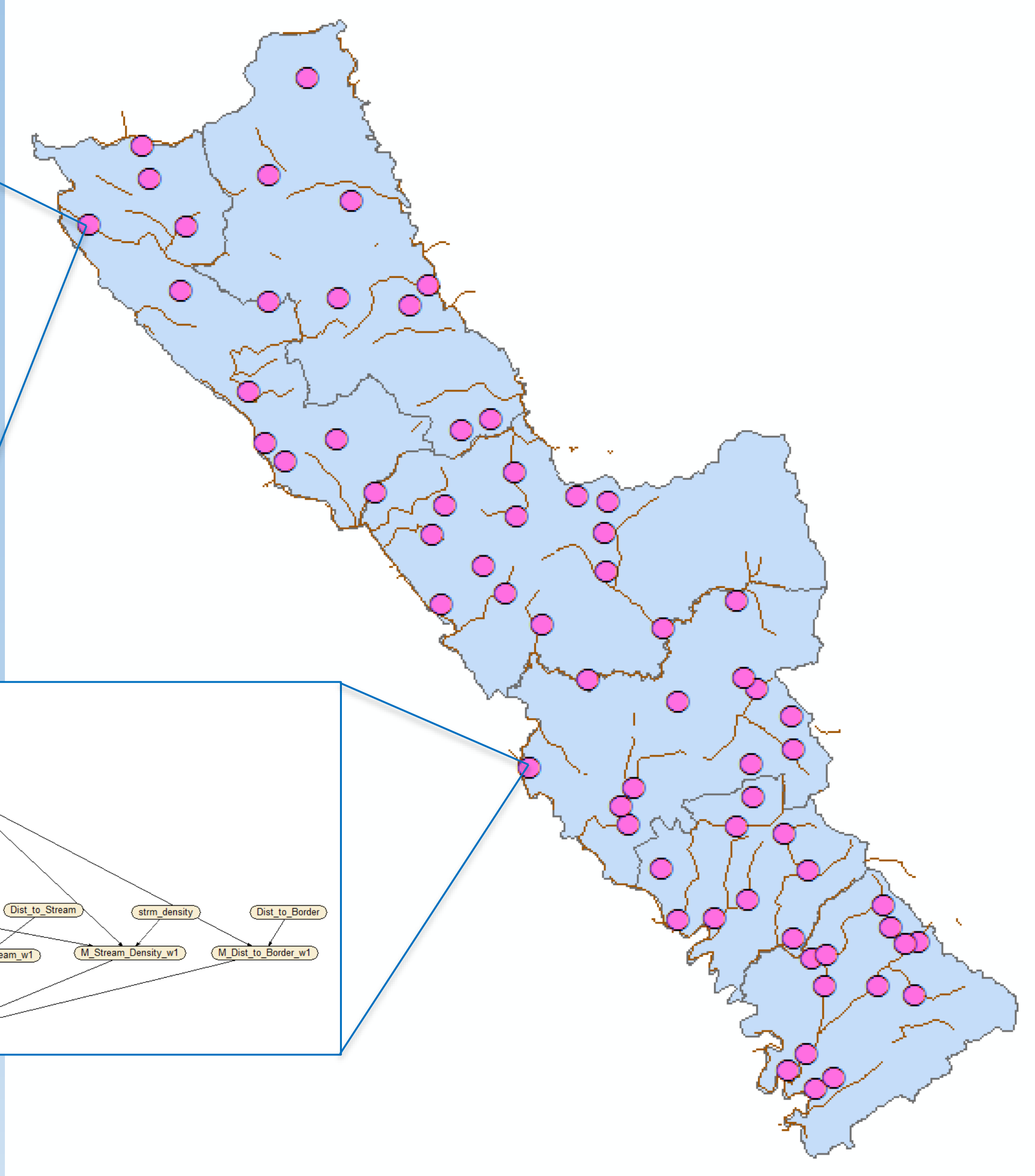


Linking Bayes Net to GIS

One Bayes net per village

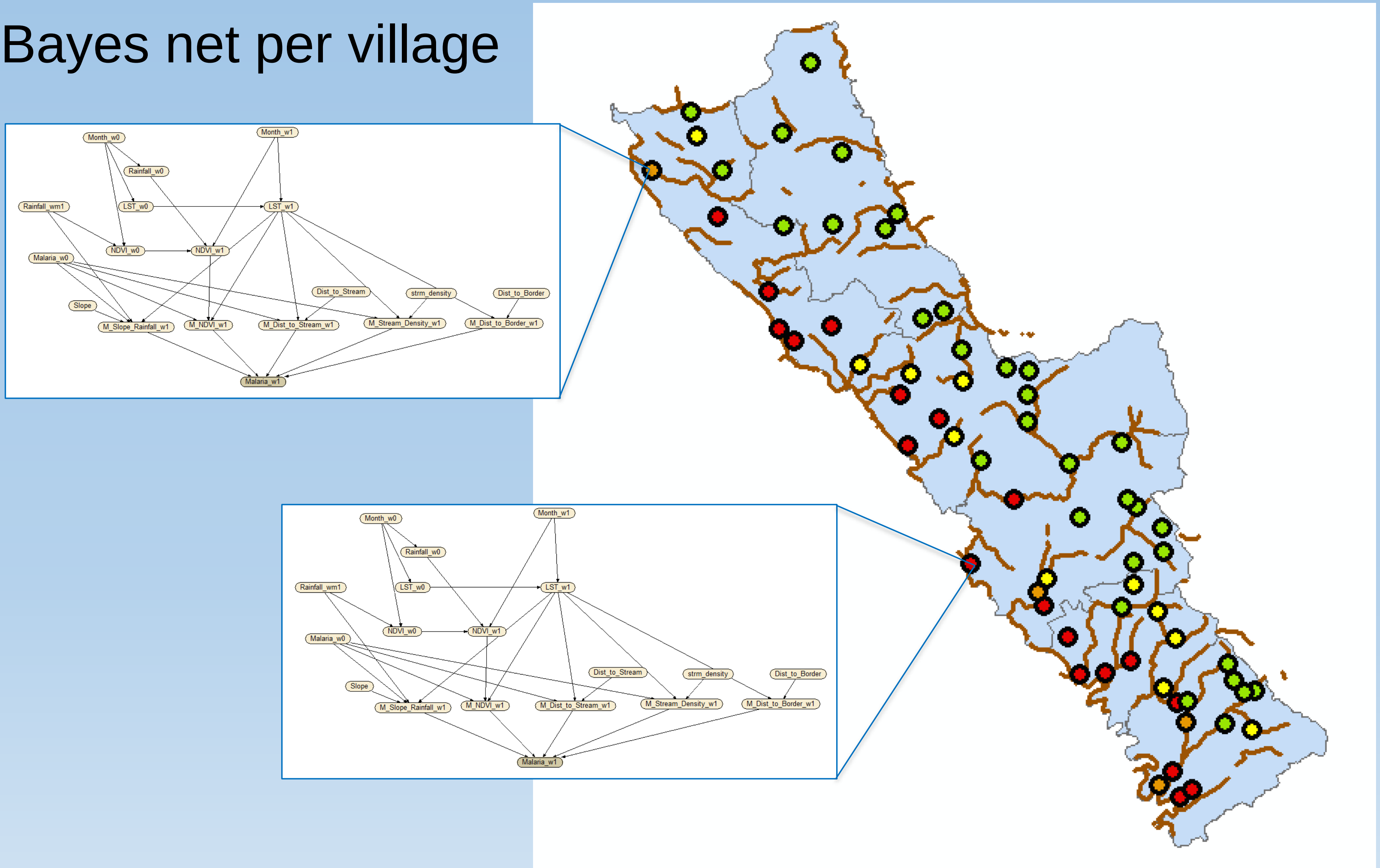






Linking Bayes Net to GIS

One Bayes net per village



Prediction Accuracy (MAE)

Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	3.387	3.404	3.454	3.504	3.43	3.457
13 med	1.886	2.073	2.129	2.319	2.299	2.411
14 low	0.305	0.368	0.422	0.463	0.519	0.538
All 66	1.415	1.501	1.557	1.644	1.657	1.729

BN outperforms ARIMA, ARIMAX, Linear Regression, Poisson Regression

- High incidence {Min: 0, Max: 82, Ave: 7.43},
- Medium incidence {Min: 0, Max: 16, Ave: 1.91}
- Low incidence {Min: 0, Max: 3, Ave: 0.099}

Comparison with ARIMA

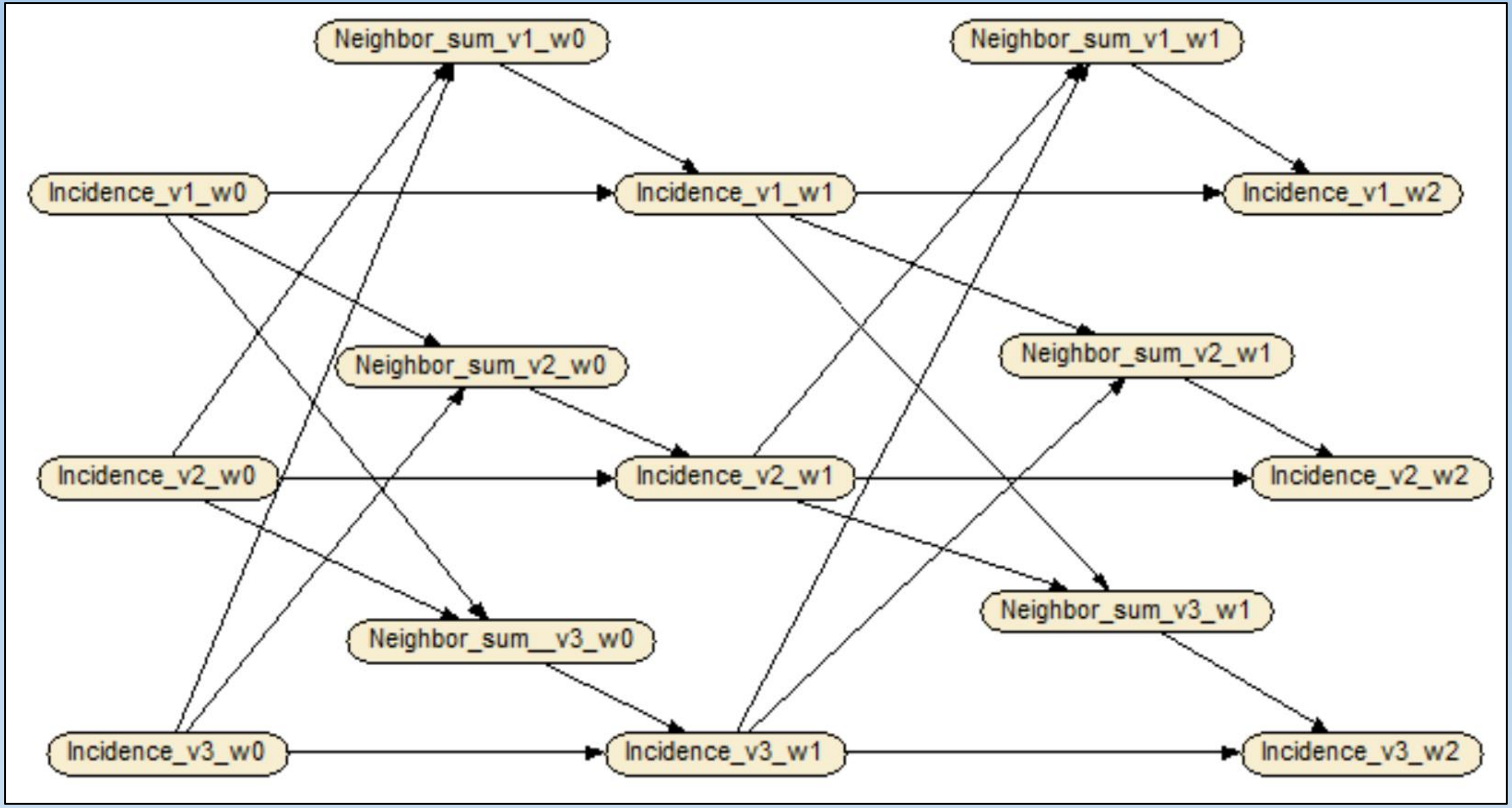
Subset	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	-3.8%	6.4%	9.3%	15.0%	20.5% *	26.9% *
13 med	-1.7%	-4.3%	-3.6%	-10.0%	-8.0%	-9.8%
14 low	-102.0% *	-145.3% *	-158.9% *	-178.9% *	-268.1% *	-304.5% *
All 66	-6.1%	-3.1%	-4.4%	-4.9%	-4.6%	-2.1%

Performance of Bayes net relative to ARIMA model.

* difference statistically significant (two-tailed T test $p < 0.05$).

Modeling Spatial Autocorrelation

- Individual village models an be linked together
- But models become too large and complex to build by hand



Solution: Knowledge-Based Model Construction

- Store model fragments in a library
- Represent as rules using probability logic

$$\forall x \, P(\mathrm{Flies}(x) \mid \mathrm{Bird}(x)) = 0.9$$

- Automatically construct models tailored to data in GIS

Model Rules

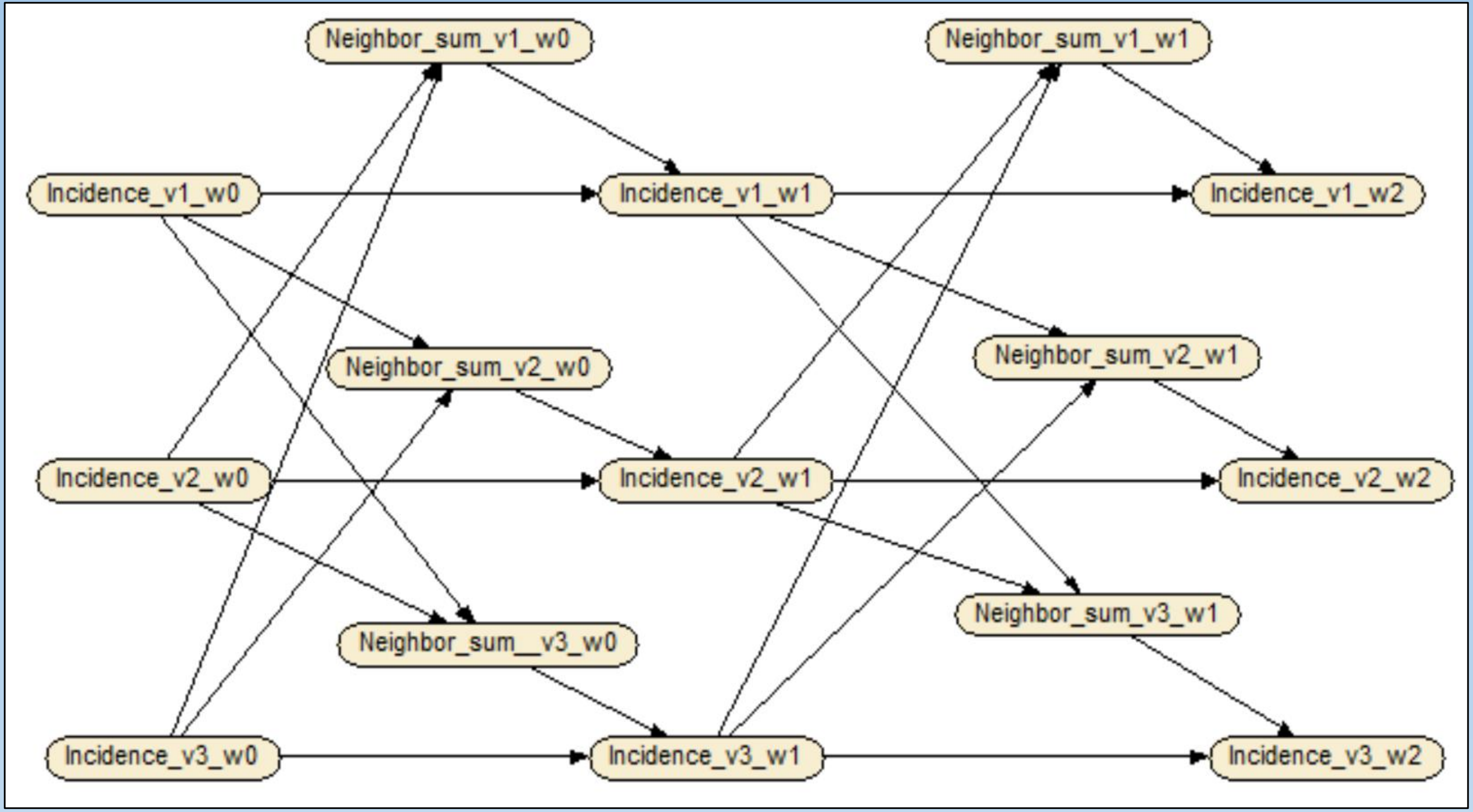
If villages are close together, malaria may spread from one to another

```
FOR(t:TIME)(x1:LOCATION)(x2:LOCATION) WHERE dist(x1,x2)<=3000 AND (x1≠x2)
        PARENT NeighborSum AT x1,t IS Cases AT x2,t
        CPT AUTO_SUM;
```

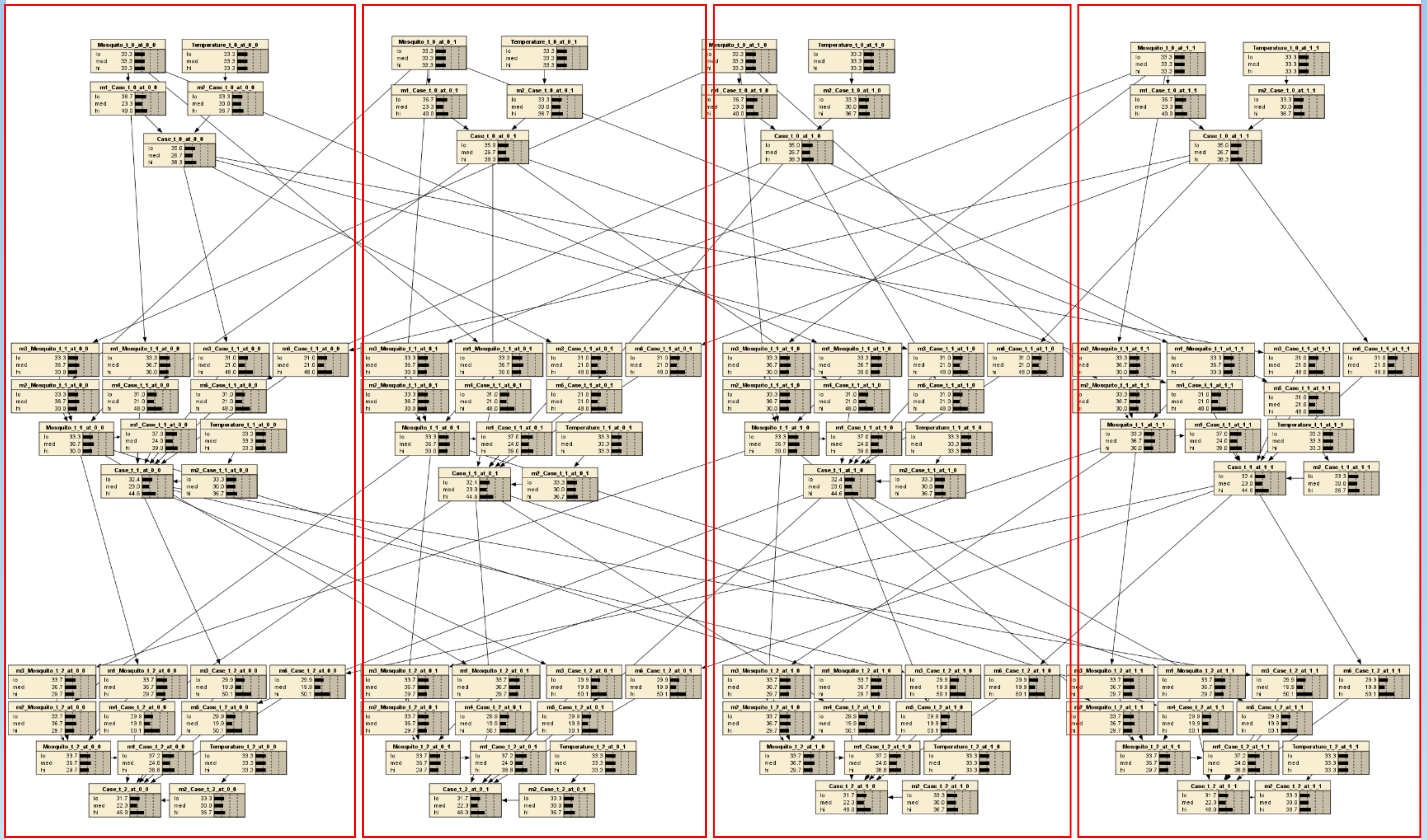
Cases has parents BorderDistance, previous week's Cases, MosquitoPop, and NearestSum

```
FOR(t1:TIME) (t2:TIME) (x:LOCATION) WHERE t2=t1+1
        PARENT Cases AT x,t2 IS BorderDistance AT x,t2 AND Cases AT x,t1 AND
        MosquitoPop AT x,t2 AND NearestSum AT x,t1
        CPT (…);
```

Group of three neighboring villages



Model for entire problem space



Improvement from Autocorrelation Model

Village No.	Total Incidence	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
109	2672	17.8% *	15.6% *	11.8%	8.92%	3.34%	5.78%
201	412	1.18%	7.75%	8.52%	4.24%	7.66%	5.38%
208	274	0.88%	3.34%	11.3% *	9.77%	10.6%	10.4%
205	83	12.9% *	22.5% *	22.2% *	18.2%	12.7%	9.05%
410	79	5.03%	4.06%	6.84%	9.15%	11.0%	8.84%
107	51	4.02%	6.17% *	12.5%	16.2%	21.6% *	20.9%

* difference statistically significant (two-tailed T test $p < 0.05$)

Ongoing and Future Work

- Evaluate on larger area
- Integrate prediction and diagnosis
- Clustering to improve prediction accuracy

- Predictive modeling of dengue
  - Incorporation of breeding site data

Thank You