新冠病毒数据分析系列公益讲座
COVID-19 Data Analysis Webinars

COVID19
Spatiotemporal Analysis and Prediction
Based on Workflow Platform
基于工作流的疫情数据时空预测模型与统计分析

Lingbo Liu Shuming Bao Taohu

武汉大学城市设计学院
SCHOOL OF URBAN DESIGN
WUHAN UNIVERSITY

China Data Lab

GRoNC Global Research on Novel Coronavirus
http://chinadatalab.org/

HARVARD
Dataverse

BACKGROUND

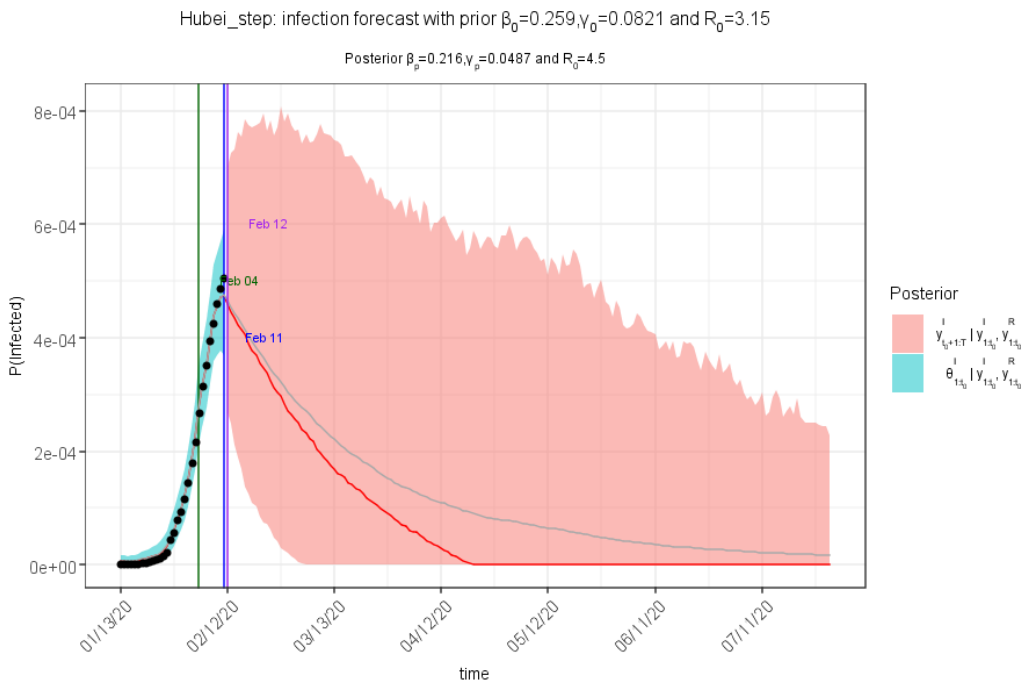
Time Varying Transmission Rate in SIR

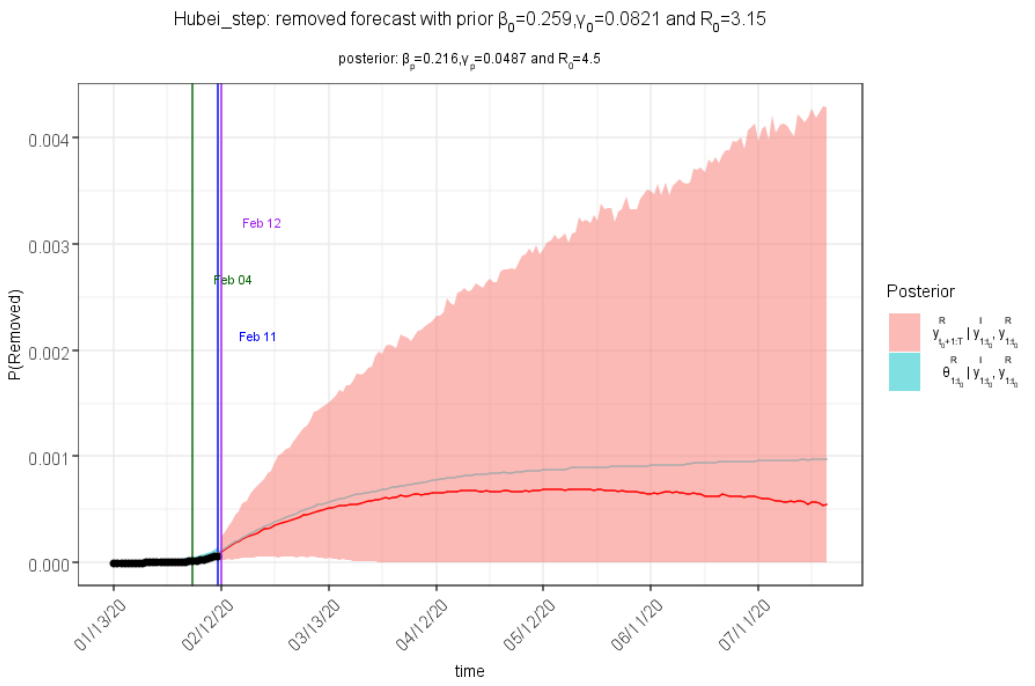
SIR model: eSIR packages in R

Song Lab
Department of Biostatistics

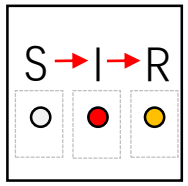
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MICHIGAN

https://github.com/lilywang1988/eSIR

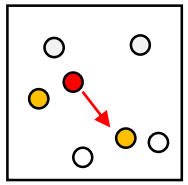




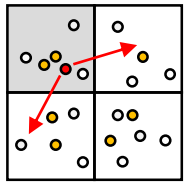
BACKGROUND



SIR Model



Spatial Contact



Neighborhood Effect

Spatial contact : the transmission of Epidemic diseases

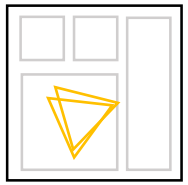
Spatial contact chance(SCC) influence the transmission rate of COVID19

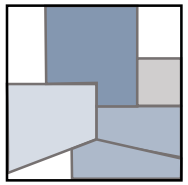
- Basic SIR model lack parameters representing the changes of spatial contact
- Basic SIR model ignore the spatial influence of adjacent areas

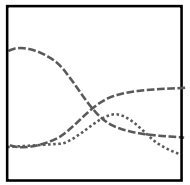
- Investigate the spatial effect of COVID19
- Integrating the parameters affecting spatial contact chance
- Building a opensource workflow for COVID19 researches

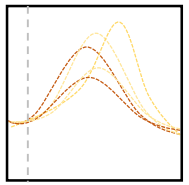
OBJECTIVES

KNIME Workflow Platform
COVID19 Spatiotemporal Analysis


KNIME


Spatial Effect


SIR Model

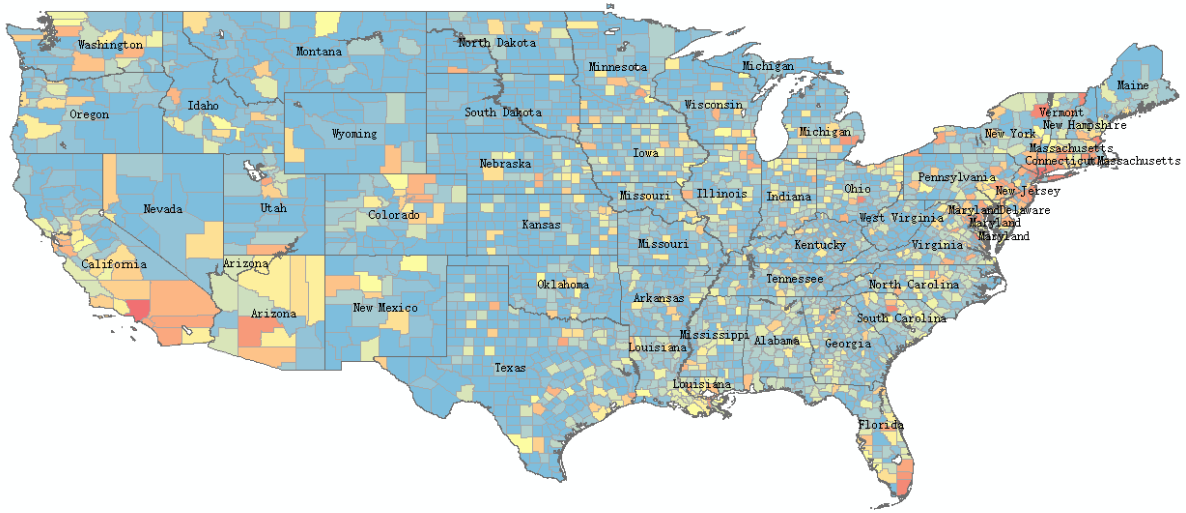

Spatial SIR

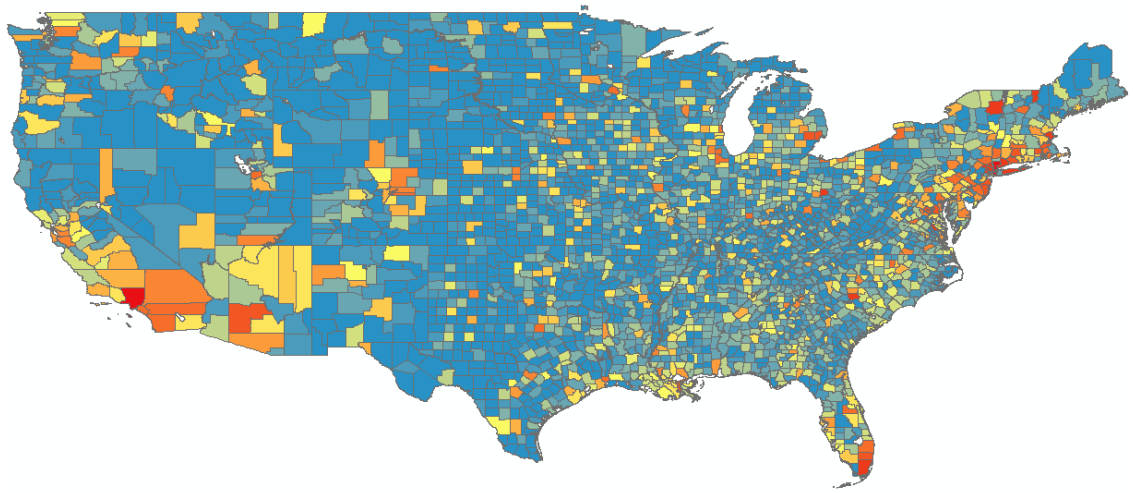
- Building COVID19 dashboard with KNIME component layout
- Spatial autocorrelation analysis with KNIME R extension
- Modelling COVID19 with SIR and Logistic Growth Model
- Spatial Panel Model integrated with SIR

DATA

- COVID19 data from nCov2019 package
- COVID19 data of Continental America (County level)

https://dataverse.harvard.edu/dataverse/cdl_dataverse

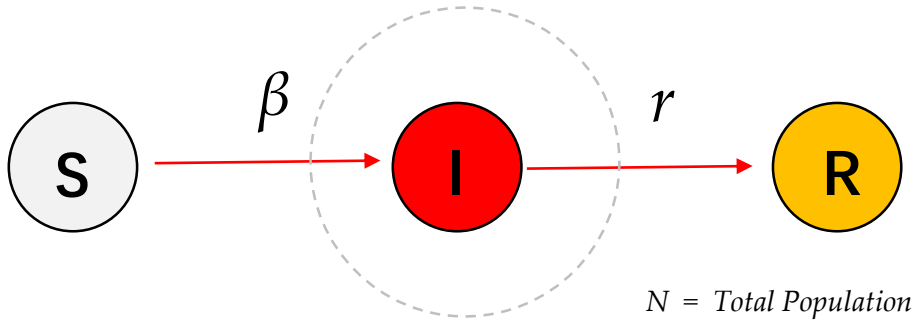




METHODOLOGY

- **Basic SIR model**

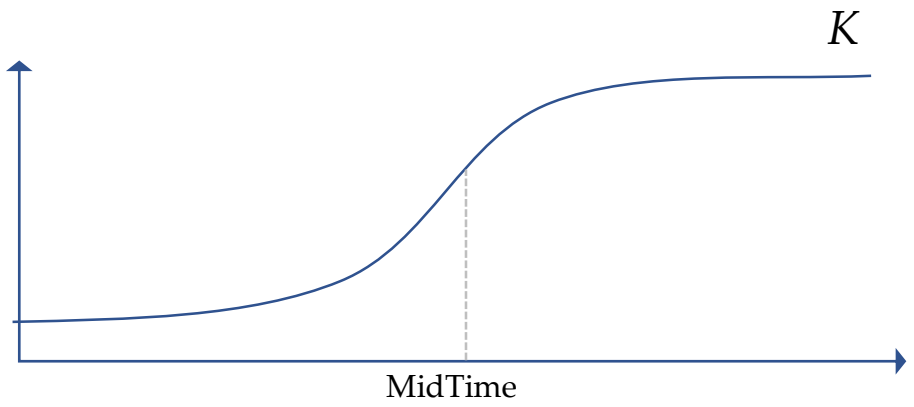
$$N = S + I + R$$
$$dS = -\beta * S * I/N$$
$$dI = \beta * S * I/N - r * I$$
$$dR = r * I$$



N = *Total Population*
S = *Susceptible population*
I = *Infected population*
R = *Removed (Recover/ Death)*

- **Logistic growth model**

$$I = K * \frac{1}{1 + e^{\frac{\text{MidTime} - t}{Scale}}}$$



METHODOLOGY

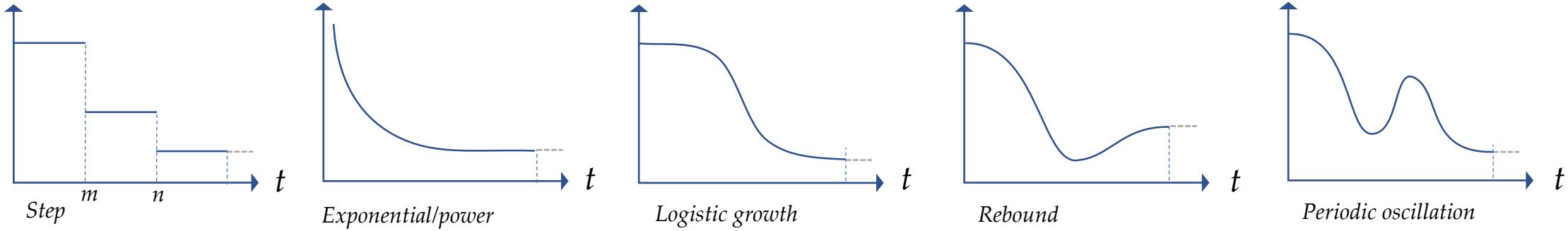
- **SIR model with time-varying SCC** $dS = -\beta SI/N$ $\beta = f(t) * \beta_0$

$$f(t) = \begin{cases} a & 0 < t < m \\ b & m < t < n \\ c & t > n \end{cases}$$

$$f(t) = K_0 * a^t$$

$$f(t) = K_0 * t^a$$

$$f(t) = K_0 * \frac{1}{1+e^{\frac{X-t}{a}}}$$

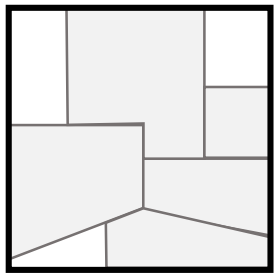


eSIR： *tvt.eSIR / qh .eSIR*

Possible function of f(t)

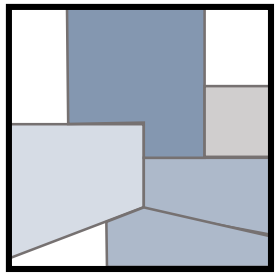
METHODOLOGY

- **Spatial Correlation**



- Moran'I

Global Spatial Statistics



- Local Moran'I

Local Spatial Statistics
LISA local Indicators of Spatial Autocorrelation

$$I(d) = \frac{\sum_{i}^{n}\sum_{j\neq i}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i}^{n}\sum_{j\neq i}^{n} w_{ij}}$$

$$I_i(d) = Z_i \sum_{j\neq i}^{n} w_{ij} Z_j$$

DATA ANALYSIS

- **SIR integrated with Spatial Correlation**

$$y = \alpha X + \varepsilon$$

Spatial Panel model $y = \rho \Sigma y + \theta Wx + \alpha X + \varepsilon \quad \varepsilon = \lambda Wx + \mu$

SLM $y = \rho \Sigma y + \alpha X + \varepsilon$

SEM $y = \alpha X + \varepsilon \quad \varepsilon = \lambda Wx + \mu$

SDM $y = \rho \Sigma y + \theta Wx + \alpha X + \varepsilon$

dI_i *(Daily new cases in Zone I)* $= \beta_i S_i I_i / N_i$

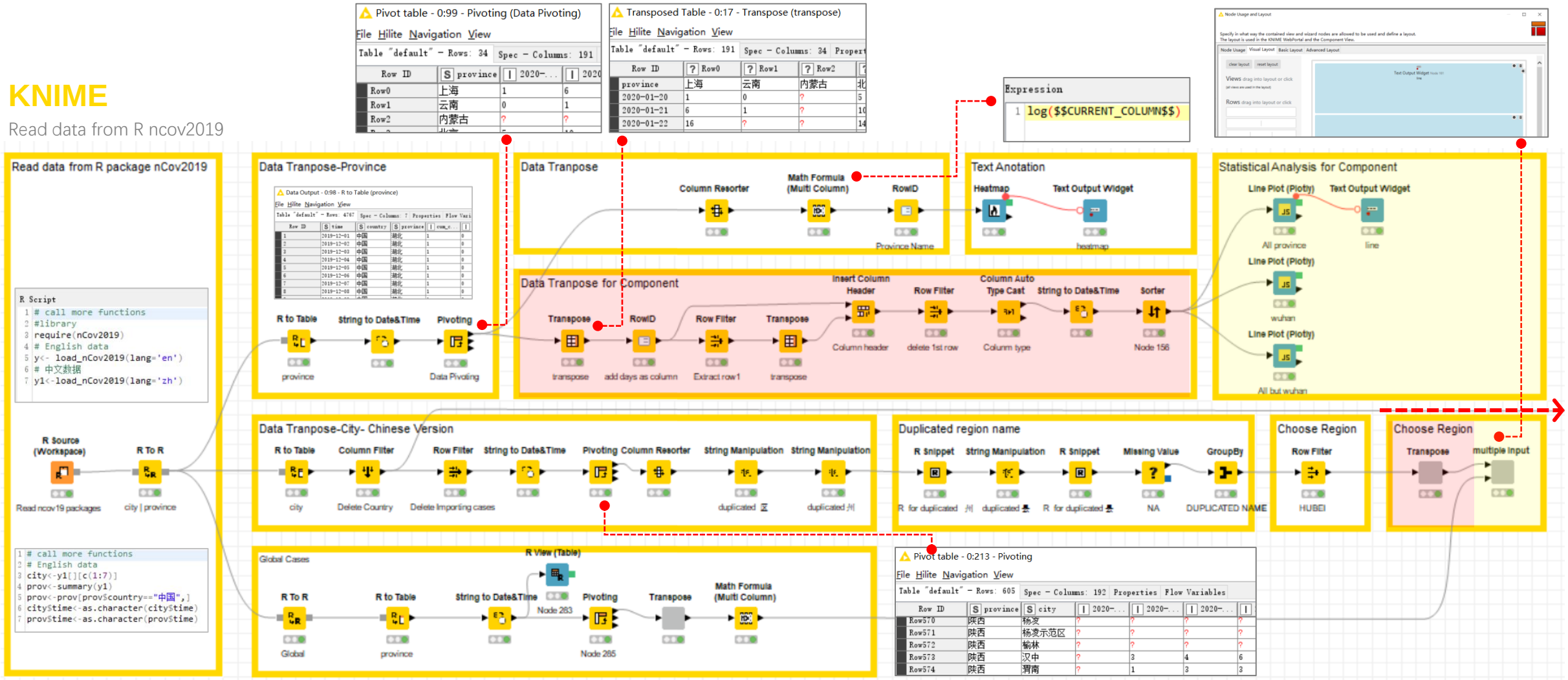
$y = \beta x_1 + r x_2$

$y = dI_i \qquad x_1 = S_i I_i / N_i$

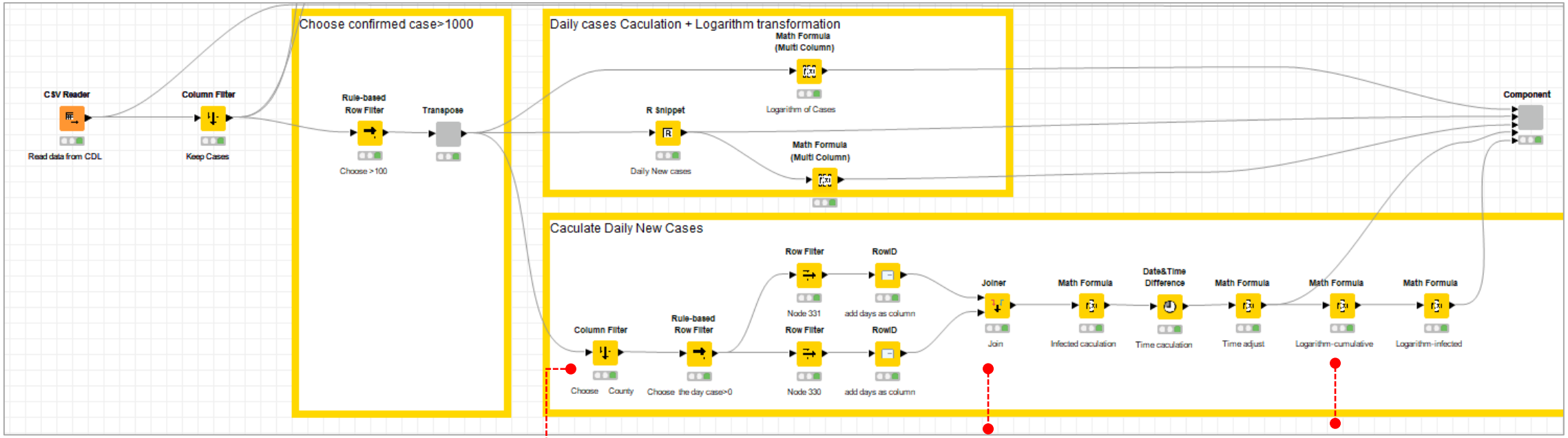
RESULTS COVID19 DASHBOARD WITH KNIME

KNIME

Read data from R ncov2019



RESULTS COVID19 DASHBOARD WITH KNIME



KNIME

Read data from CSC reader

Filtered - 0:328 - Rule-based Row Filter

File Hilite Navigation View

Table "default" – Rows: 60 Spec – Columns

Row ID	Milwa...	day
Row52	2	2020-03-13
Row53	6	2020-03-14
Row54	7	2020-03-15
Row55	13	2020-03-16
Row56	24	2020-03-17

Row ID	Milwa...	day	Milwa...	day (#1)
Row0	6	2020-03-14	2	2020-03-13
Row1	7	2020-03-15	6	2020-03-14
Row2	13	2020-03-16	7	2020-03-15
Row3	24	2020-03-17	13	2020-03-16
Row4	47	2020-03-18	24	2020-03-17
Row5	62	2020-03-19	47	2020-03-18
Row6	85	2020-03-20	62	2020-03-19
Row7	126	2020-03-21	85	2020-03-20
Row8	182	2020-03-22	126	2020-03-21
Row9	204	2020-03-23	182	2020-03-22

log(x)

Expression

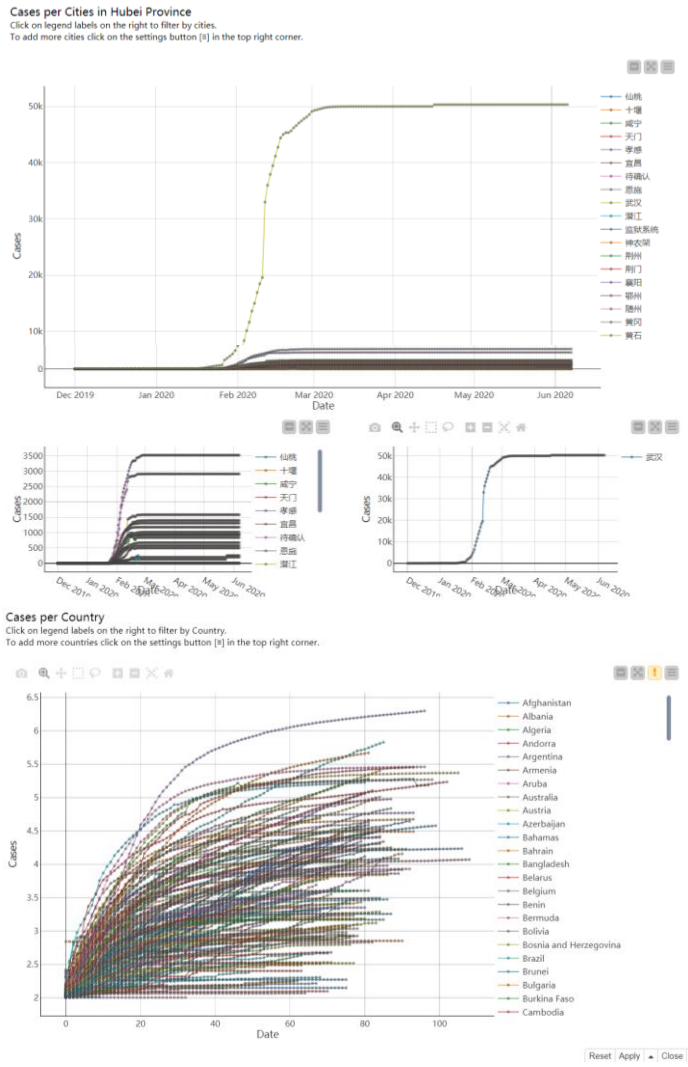
```
ln($Milwaukee County$)
```

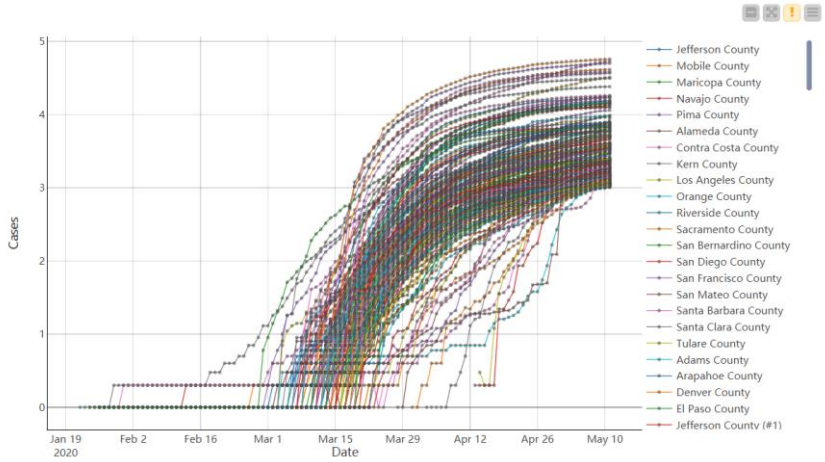
Append Column:

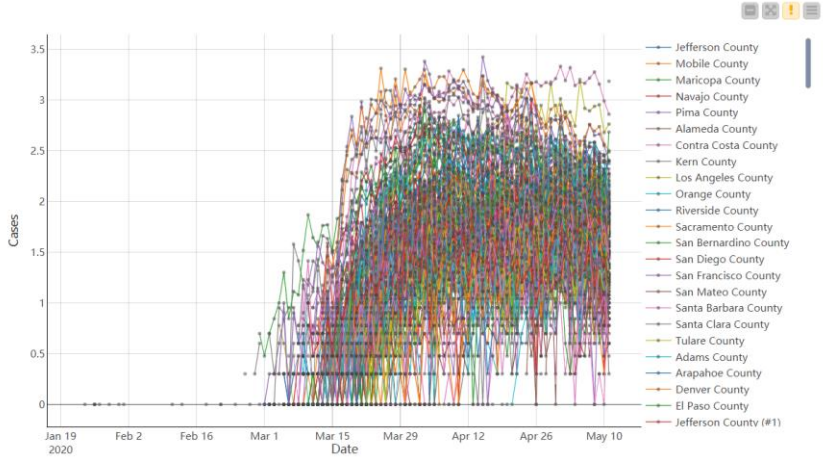
Replace Column: Milwaukee County

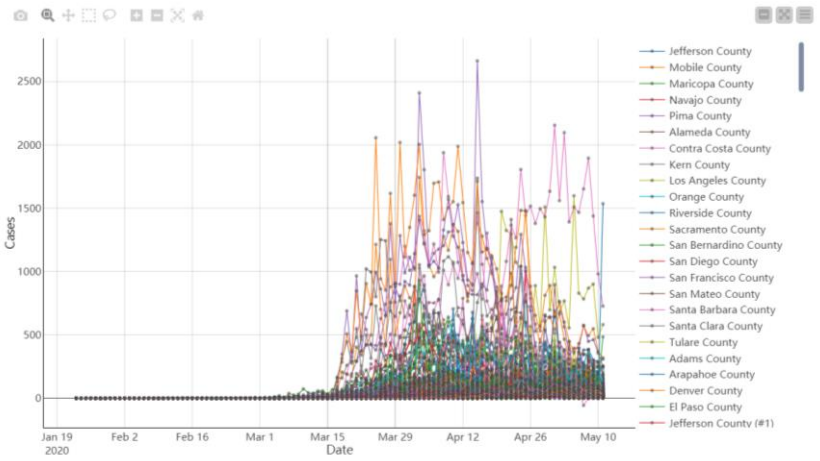
Convert to Int

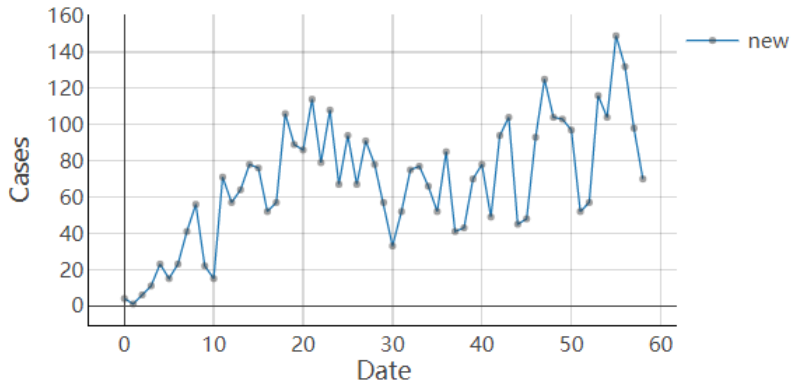
RESULTS COVID19 DASHBOARD WITH KNIME



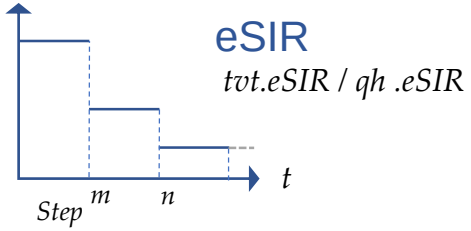




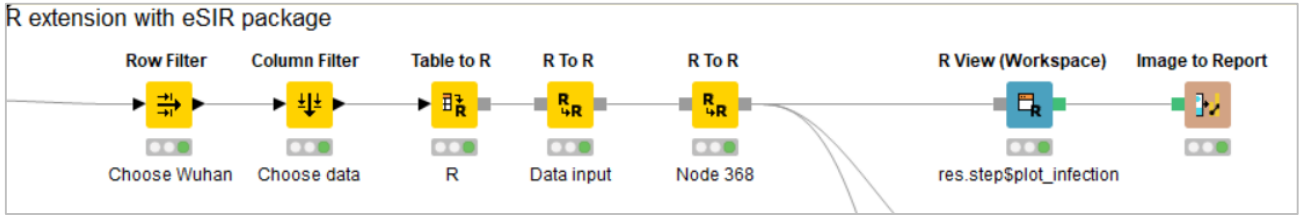


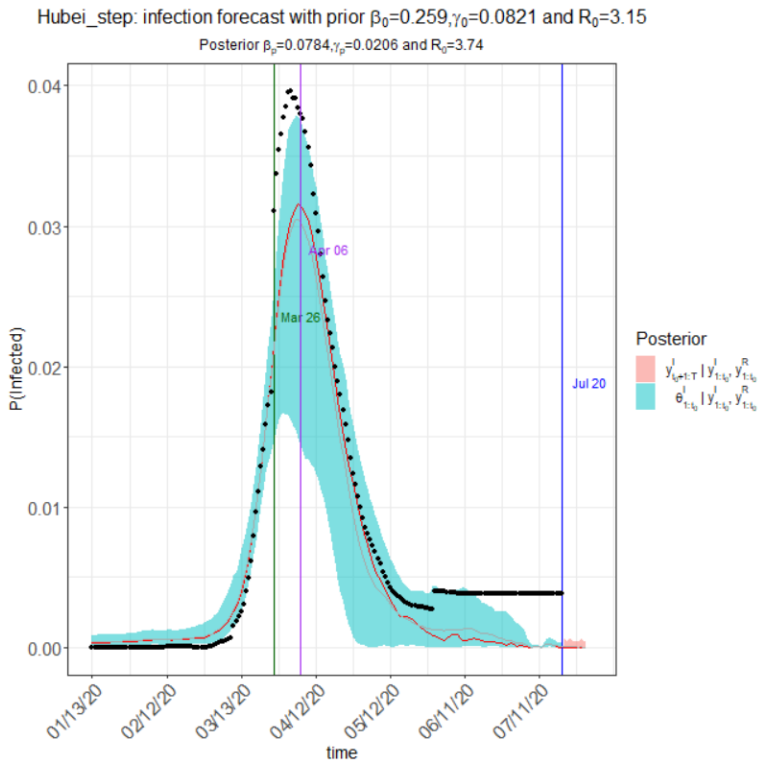


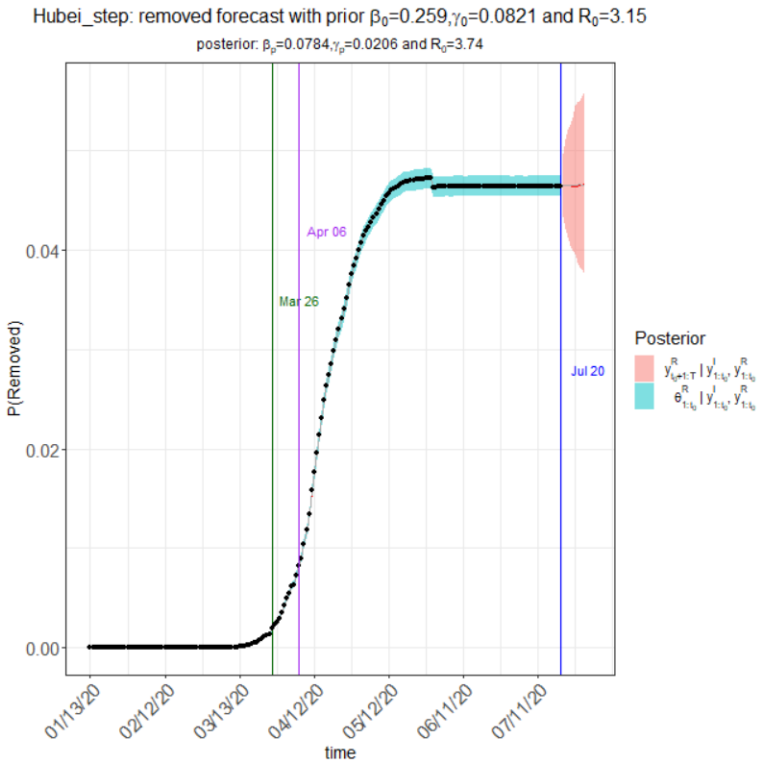
RESULTS eSIR MODELLING

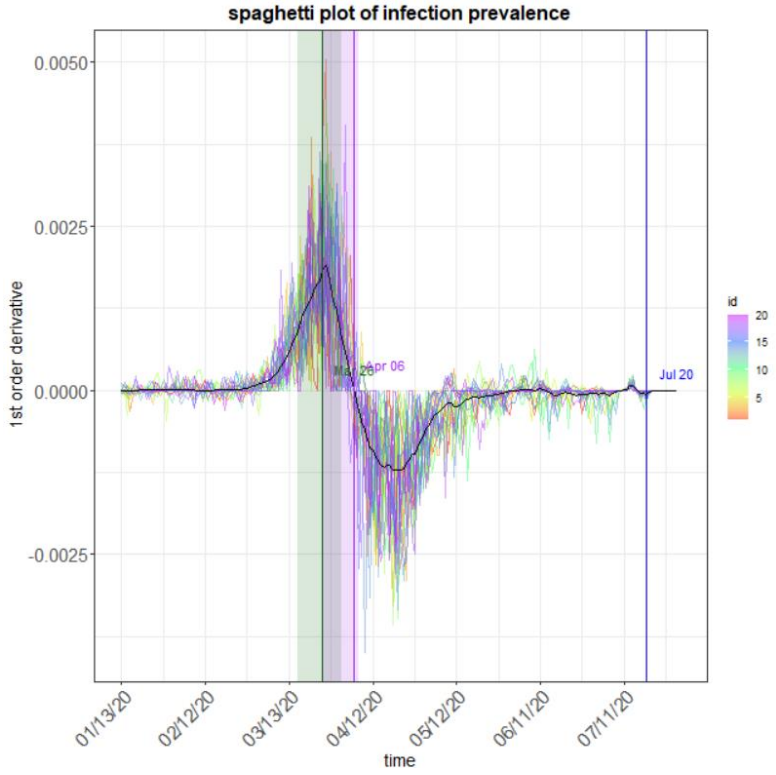


KNIME



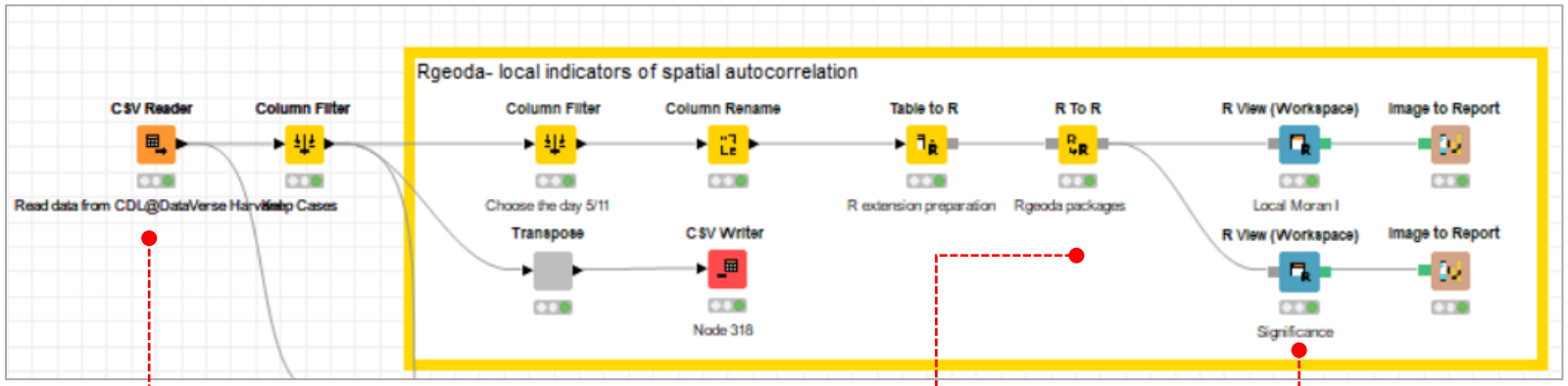


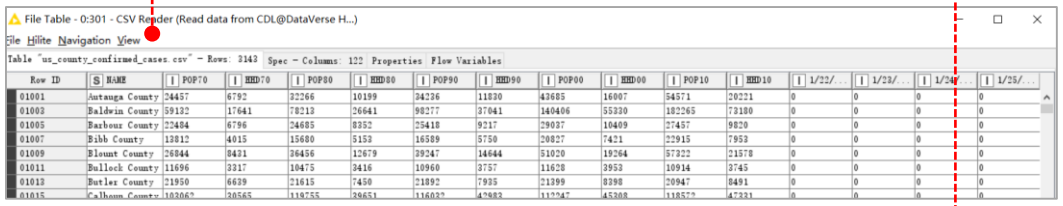




RESULTS

LOCAL MORAN'S I OF CONFIRMED CASES





R Script

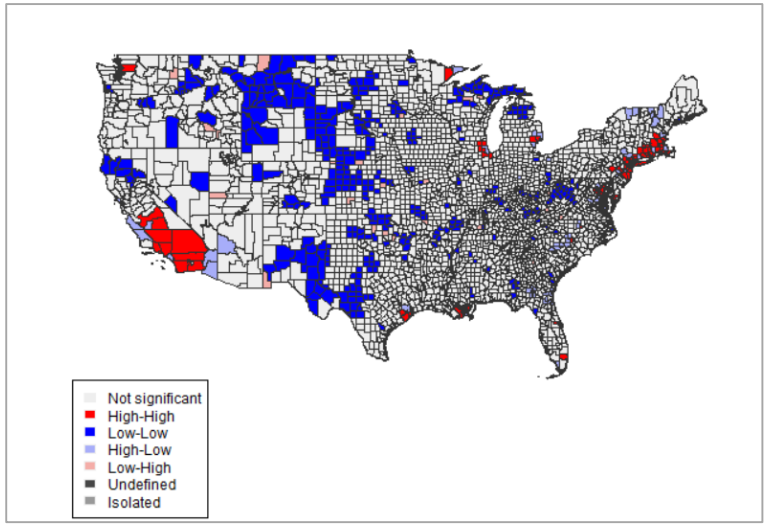
```
# call more functions
library(rgdal)
library(rgeoda)
library(sf)

county_sf <- st_read("D:/CDL/dataverse/UScounty/COUNTY_MAP/countytest.shp")
county_sf<-merge(county_sf,df,all.x=TRUE)
county <- sf_to_geoda(county_sf, with_table = TRUE)
county0w <- queen_weights(county)

Confirm <- county_sf$Confirm
lisa <- local_moran(county0w, Confirm)
lisa_colors <- lisa$GetColors()
lisa_labels <- lisa$GetLabels()
lisa_clusters <- lisa$GetClusterIndicators()
lisa_p <- lisa$GetPValues()
```

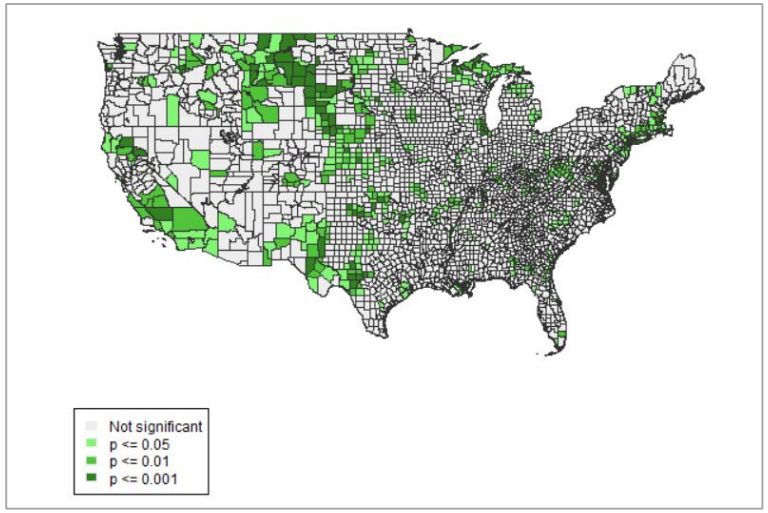
R Script

```
plot(st_geometry(county_sf),
     col=sapply(lisa_clusters, function(
     border = "#333333", lwd=0.2)
title(main = "Local Moran Map of COVID19
legend('bottomleft', legend = lisa_label
```

R Script

```
p_labels <- c("Not significant", "
p_colors <- c("#eeeeee", "#84f576"
plot(st_geometry(county_sf),
     col=sapply(lisa_p, function(x
       if (x <= 0.001) return(p_co
       else if (x <= 0.01) return(
       else if (x <= 0.05) return
       else return(p_colors[1])
       }),
     border = "#333333", lwd=0.2)
```

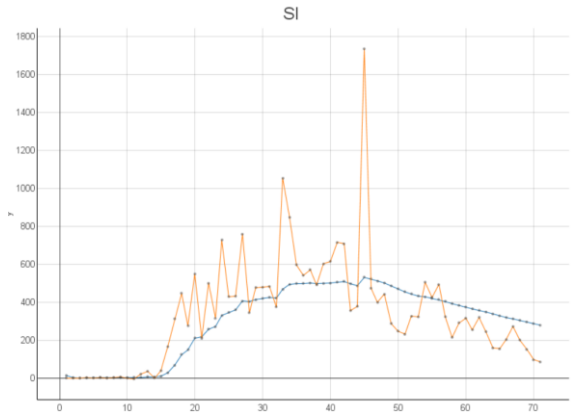


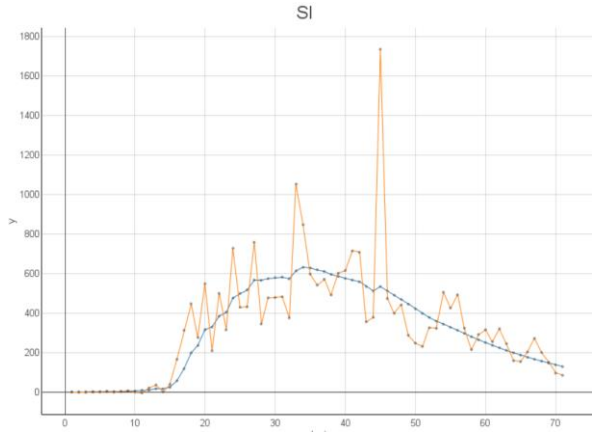


RESULTS SIR WITH TIME VARYING SCC

$$dS = \beta SI/N \quad \beta = f(t)^* \beta_0$$

$$Ln(dS) = Ln\beta_0 + ln(SI/N) + ln\, f(t)$$





Linear Re...

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
INS	0.635	0.0386	16.4442	0.0
Intercept	-0.0425	0.3208	-0.1326	0.8949

Multiple R-Squared: 0.7967
Adjusted R-Squared: 0.7938

$$f(t) = \beta_0$$

Linear Re...

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
dayt	-2.492	0.4017	-6.204	3.73E-8
INS	1.3568	0.1204	11.2666	0.0
Intercept	2.6083	0.4992	5.2246	1.81E-6

Multiple R-Squared: 0.8702
Adjusted R-Squared: 0.8664

$$f(t) = \beta_0 * t^{a}$$

Linear Regr...

File

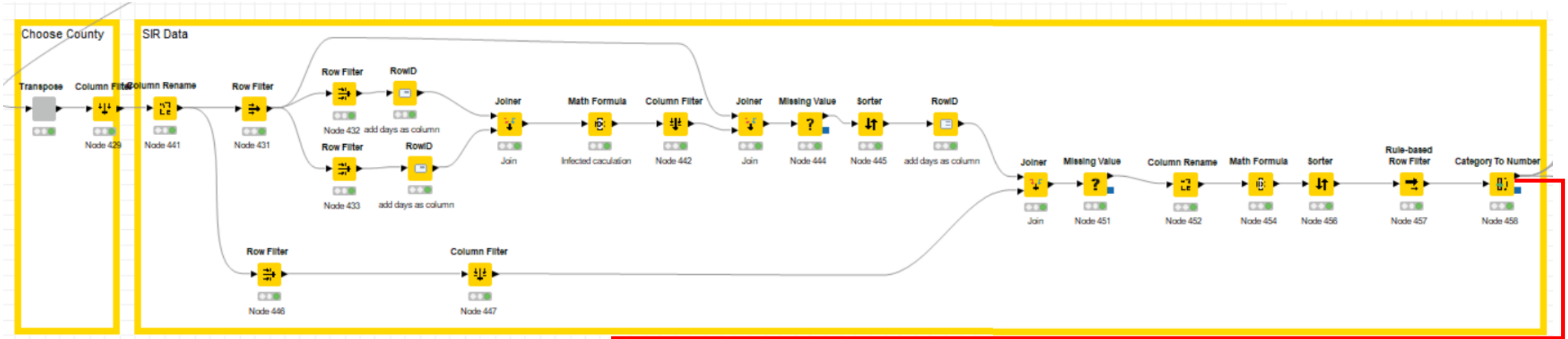
Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
dayt	-0.0691	0.0074	-9.3095	9.35E-14
INS	1.0384	0.0504	20.5921	0.0
Intercept	-0.6779	0.2249	-3.0146	0.0036

Multiple R-Squared: 0.9106
Adjusted R-Squared: 0.908

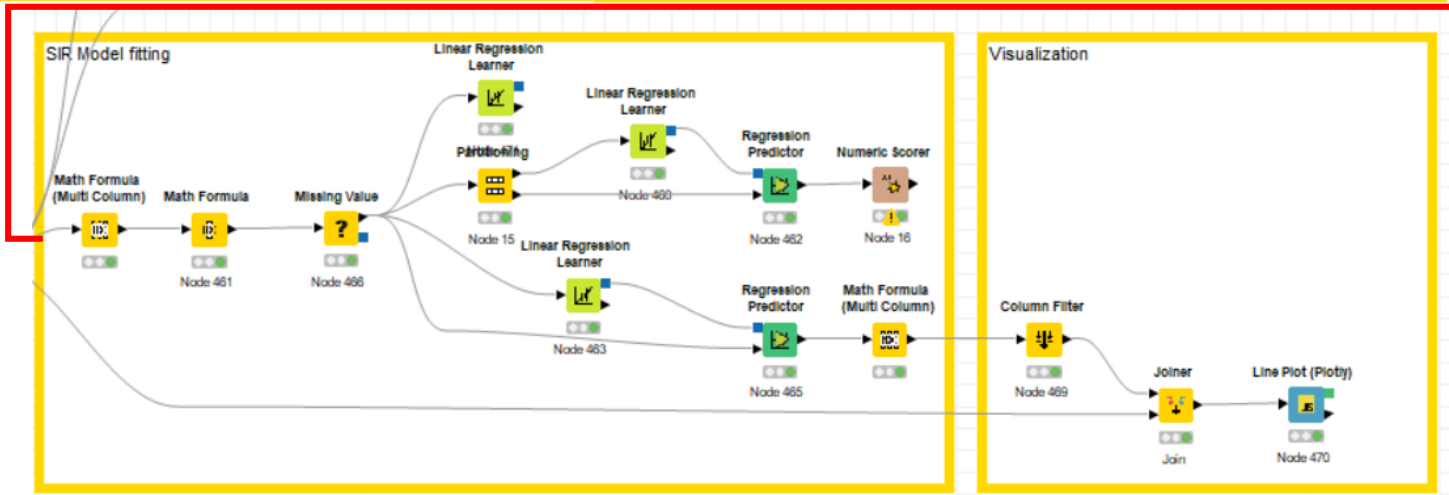
$$f(t) = \beta_0 * a^{t}$$

RESULTS SIR WITH TIME VARYING SCC

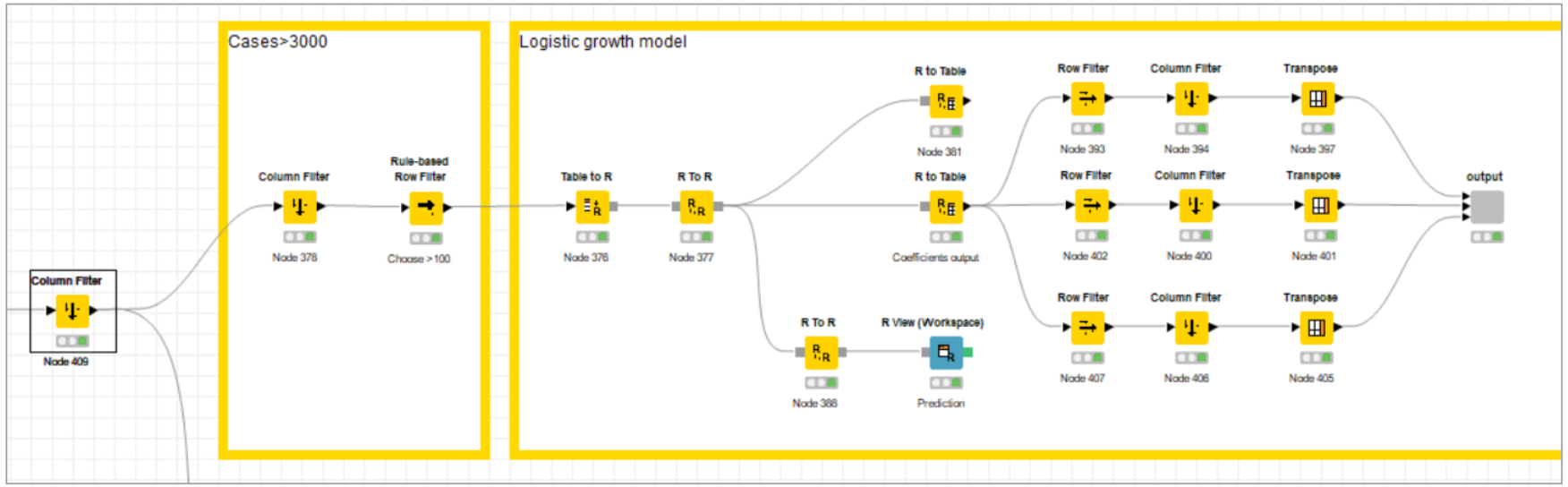


Queen...	Richm...	Rockl...	Schen...	day
80117	165516	99242	62886	1/10
0	0	0	0	1/22/2020
0	0	0	0	1/23/2020
0	0	0	0	1/24/2020
0	0	0	0	1/25/2020
0	0	0	0	1/26/2020
0	0	0	0	1/27/2020
0	0	0	0	1/28/2020
0	0	0	0	1/29/2020
0	0	0	0	1/30/2020
0	0	0	0	1/31/2020
0	0	0	0	2/1/2020
0	0	0	0	2/2/2020

Row ID	County	day	new
Row70_Row67	6060	3/30/2020	478
Row71_Row68	6539	3/31/2020	479
Row72_Row69	7022	4/1/2020	483
Row73_Row70	7398	4/2/2020	376
Row74_Row71	8452	4/3/2020	1,054
Row75_Row72	9300	4/4/2020	848
Row76_Row73	9898	4/5/2020	598
Row77_Row74	10440	4/6/2020	542
Row78_Row75	11012	4/7/2020	572
Row79_Row76	11504	4/8/2020	492
Row80_Row77	12106	4/9/2020	602
Row81_Row78	12722	4/10/2020	616
Row82_Row79	13437	4/11/2020	715
Row83_Row80	14145	4/12/2020	708



RESULTS LOGISTIC GROWTH MODEL

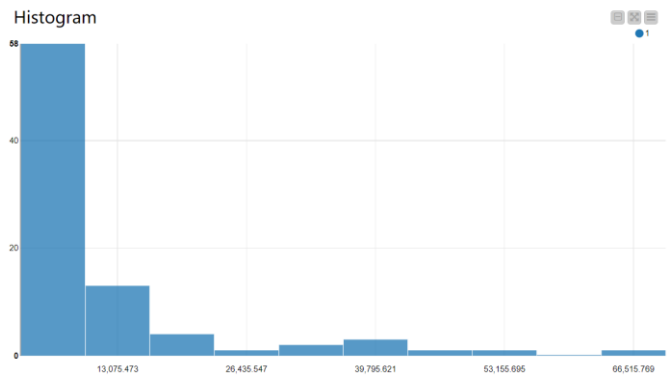


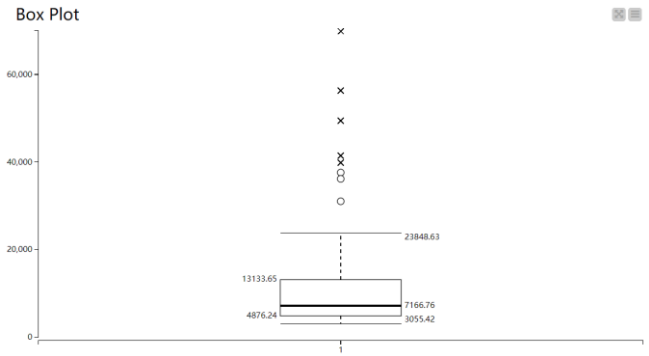
time	Maric...	t1	Los A...	t2	Orang...	t3	River...	t4	San B...	t5	
1	5.17	5	8.042	5	3.925	5	31.24	47	64.576	54	1
2	5.589	6	8.828	6	4.236	6	35.252	48	71.012	55	1
3	6.043	7	9.692	7	4.572	7	39.775	49	78.075	56	2
4	6.533	8	10.641	8	4.935	8	44.874	50	85.821	57	2
5	7.062	9	11.682	9	5.326	9	50.62	51	94.314	58	2
6	7.635	10	12.825	10	5.748	10	57.093	52	103.621	59	2
7	8.254	11	14.08	11	6.203	11	64.384	53	113.814	60	2

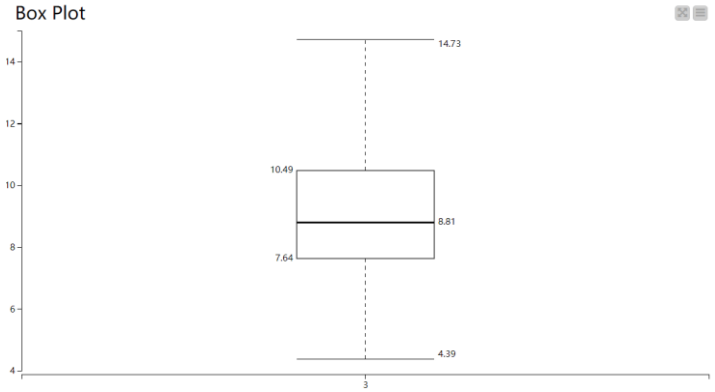
R Script

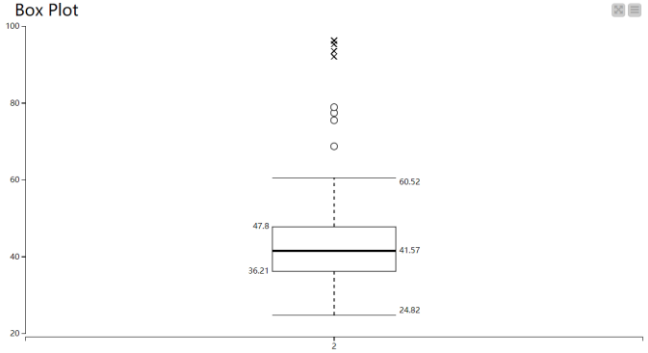
```r
# delete
dataf<-dataf[-41,]
library(broom)
library(tidyr)
library(dplyr)
tt<-300
x2<-1:tt
data0<-data.frame(time=x2)
result<-data.frame(term=c("a","b","c","d"))
nm<-length(dataf[,1])
for (i in 1:nm){
     n<-i
     x<-dataf[n,]
     xt<-t(x)
     xt<-xt[-1,]
     xt<-subset(xt,xt!=0)
     xt<-as.data.frame(xt)
     IDL<-length(xt[,1])
     xt$ID<-1:IDL
     T0<-length(x)-IDL-1
     xt$xt<-as.numeric(as.character(xt$xt))
     lm1 <- nls(xt ~ SSlogis(ID, a, b,c),data=xt)
     y2<-predict(lm1,newdata = data.frame(ID= x2))
     newdata<-data.frame(y=y2,time=x2)
     Name<-x$NAME[1]
     names(newdata)[1]<-as.character(Name)
     newdata$t<-newdata$time+T0
     data0<-merge(data0,newdata,by="time")
     res<-tidy(lm1)
     names(res)[5]<-paste(Name,sep="")
     result<-merge(result,res,by="term")
}
```

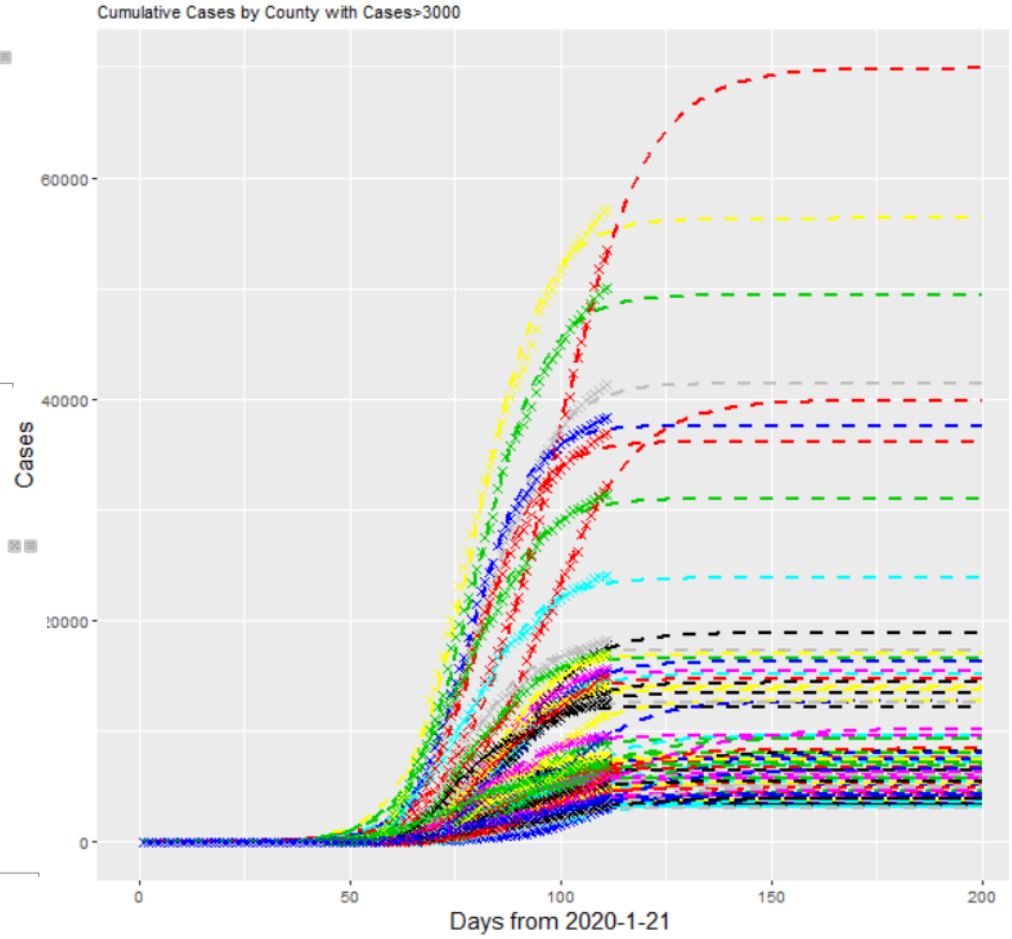
RESULTS LOGISTIC GROWTH MODEL



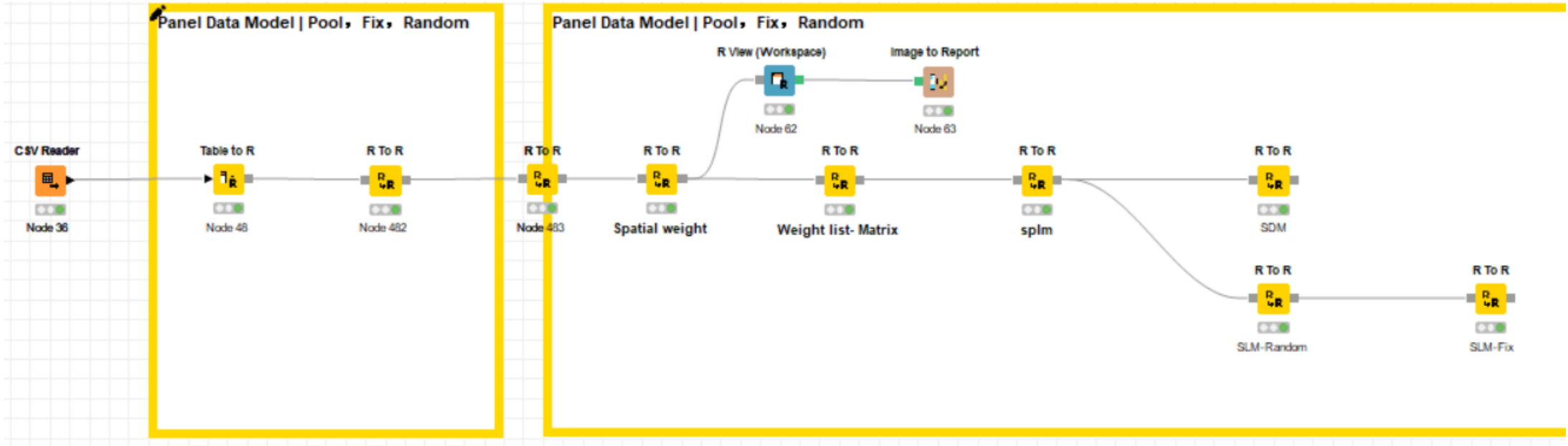








RESULTS SIR INTEGRATED WITH SPATIAL PANEL MODEL



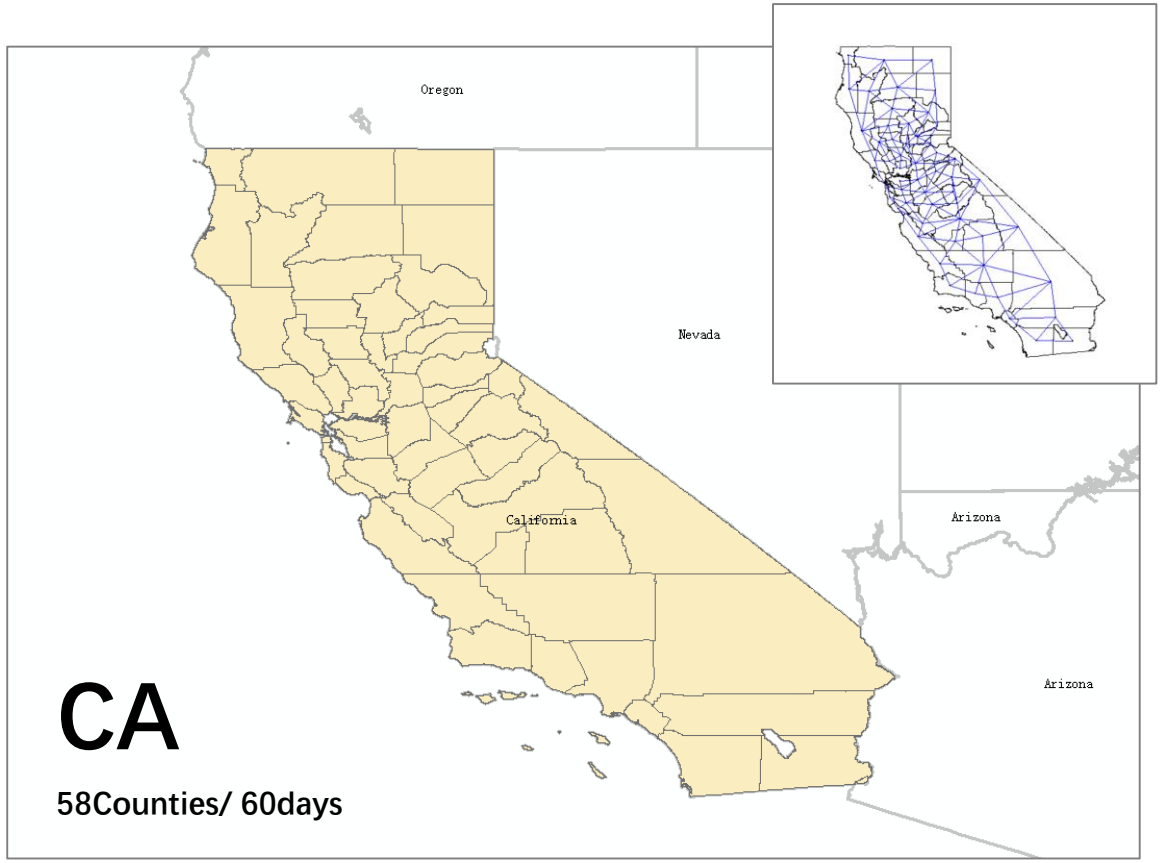
R Script

```r
# call more functions
df2<-df[df$ds!=0,]
df2<-df2[df2$SIN!=0,]
f2<-log(ds)~log(SIN)+t


pool3 = plm(f2, data=df2, model="pooling")
summary(pool3)
```

R Script

```r
#使用poly2nb方法，定义空间关系，现在共点共边（queen =T）
w_cn <- poly2nb(mdata, queen=T)
#计算出来的nb对象，还仅仅是一个临近要素列表，还并非空间权重矩阵，所
w_cn_mat<- nb2listw(w_cn, style="W", zero.policy=TRUE)
```

RESULTS SIR INTEGRATED WITH SPATIAL PANEL MODEL



PANEL DATA MODEL

dI ~ SI/N+t+N

Oneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)

Call:
plm(formula = f, data = dfr, model = "random")

Balanced Panel: n = 58, T = 60, N = 3480

Coefficients:

	Estimate	Std. Error	z-value	Pr(>|z|)	
(Intercept)	-4.4135e+00	1.5839e+00	-2.7864	0.005329	**
SIN	2.6766e-02	5.2925e-04	50.5738	< 2.2e-16	***
t	-4.4802e-02	4.3677e-02	-1.0258	0.305007	
N	2.1676e-05	7.4493e-07	29.0979	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 27738000
Residual Sum of Squares: 6488300
R-Squared: 0.76608
Adj. R-Squared: 0.76588
Chisq: 11384.1 on 3 DF, p-value: < 2.22e-16

log(dI)~log(SIN)+t

Pooling Model

Call:
plm(formula = f2, data = dfr2, model = "pooling")

Unbalanced Panel: n = 53, T = 1-60, N = 1732

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-0.3995454	0.0570217	-7.0069	3.478e-12	***
log(SIN)	0.7866188	0.0124075	63.3986	< 2.2e-16	***
t	-0.0389873	0.0015528	-25.1081	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 4387.7
Residual Sum of Squares: 1282.8
R-Squared: 0.70764
Adj. R-Squared: 0.7073
F-statistic: 2092.42 on 2 and 1729 DF, p-value: < 2.22e-16

RESULTS SIR INTEGRATED WITH SPATIAL PANEL MODEL

$$y = \rho \Sigma y + \theta Wx + \alpha X + \varepsilon \qquad \varepsilon = \lambda Wx + \mu$$

SPATIAL PANEL MODEL

Spatial panel fixed effects GM model

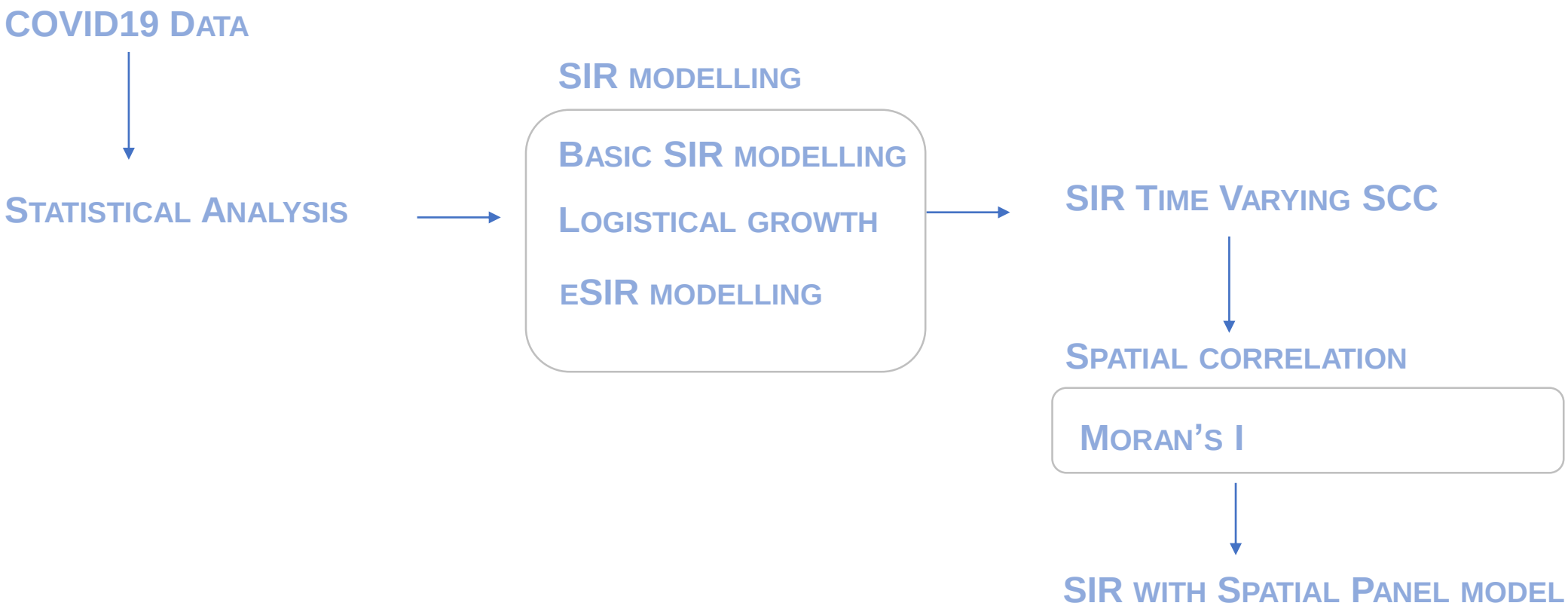
```
Call:
spgm(formula = f, data = dfr, listw = mat2listw(cw), spatial.error = TRUE,
    moments = "fullweights")

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-555.45933   -1.77061   -0.10872    0.60889  903.34542

Estimated spatial coefficient, variance components and theta:
            Estimate
rho        8.3095e-02
sigma^2_v 1.7724e+03

Coefficients:
       Estimate  Std. Error t-value Pr(>|t|)
SIN  0.02479794  0.00053399 46.4387   <2e-16 ***
t   -0.00252312  0.04582534 -0.0551   0.9561
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SDM

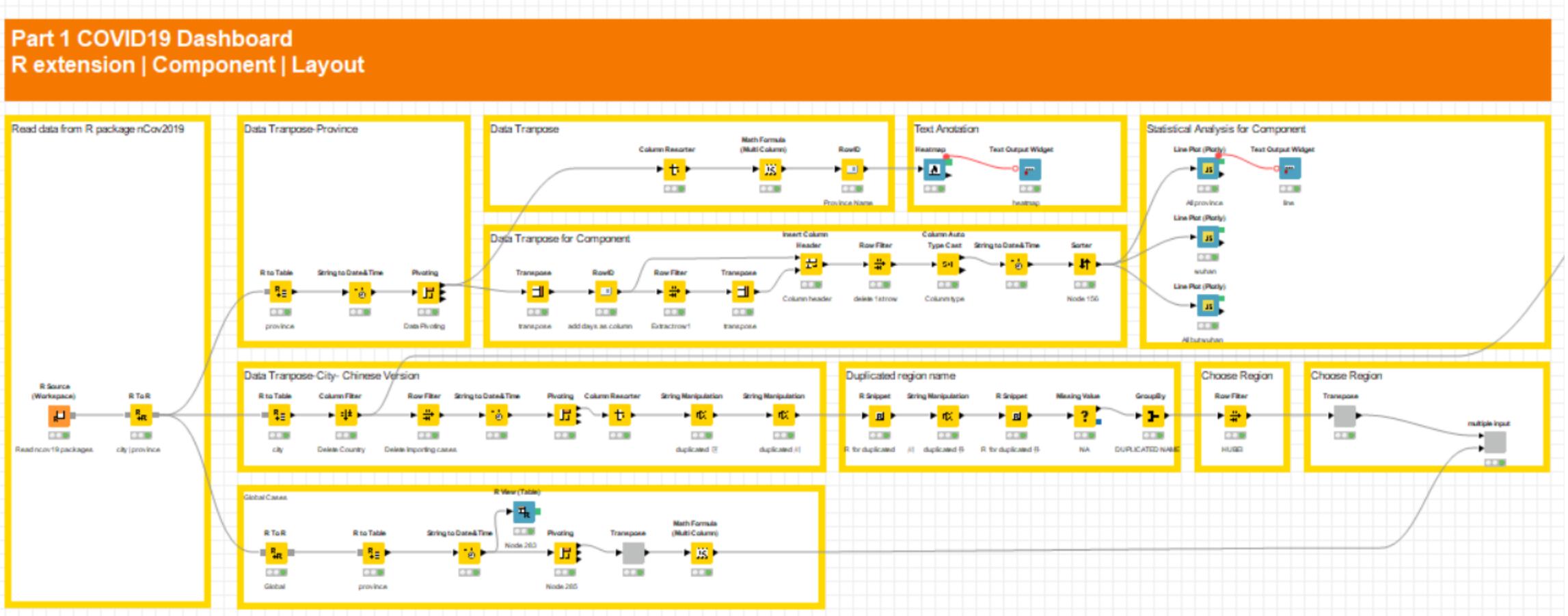
ML panel with spatial lag, AR(1) serial correlation

```
Call:
spreml(formula = f, data = dfr, w = cw, lag = TRUE, errors = "sr",
    method = "BFGS")

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-628.57   -6.70   -3.01    1.11    0.81 1033.17

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
psi 0.313712   0.016608  18.889 < 2.2e-16 ***

Spatial autoregressive coefficient:
       Estimate Std. Error t-value Pr(>|t|)
lambda 0.050831   0.017212  2.9533 0.003144 **

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  9.36792058  1.89050213  4.9553 7.224e-07 ***
SIN          0.03573586  0.00051619 69.2301 < 2.2e-16 ***
t           -0.20519461  0.05150145 -3.9842 6.769e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Flowchart for Data Analysis

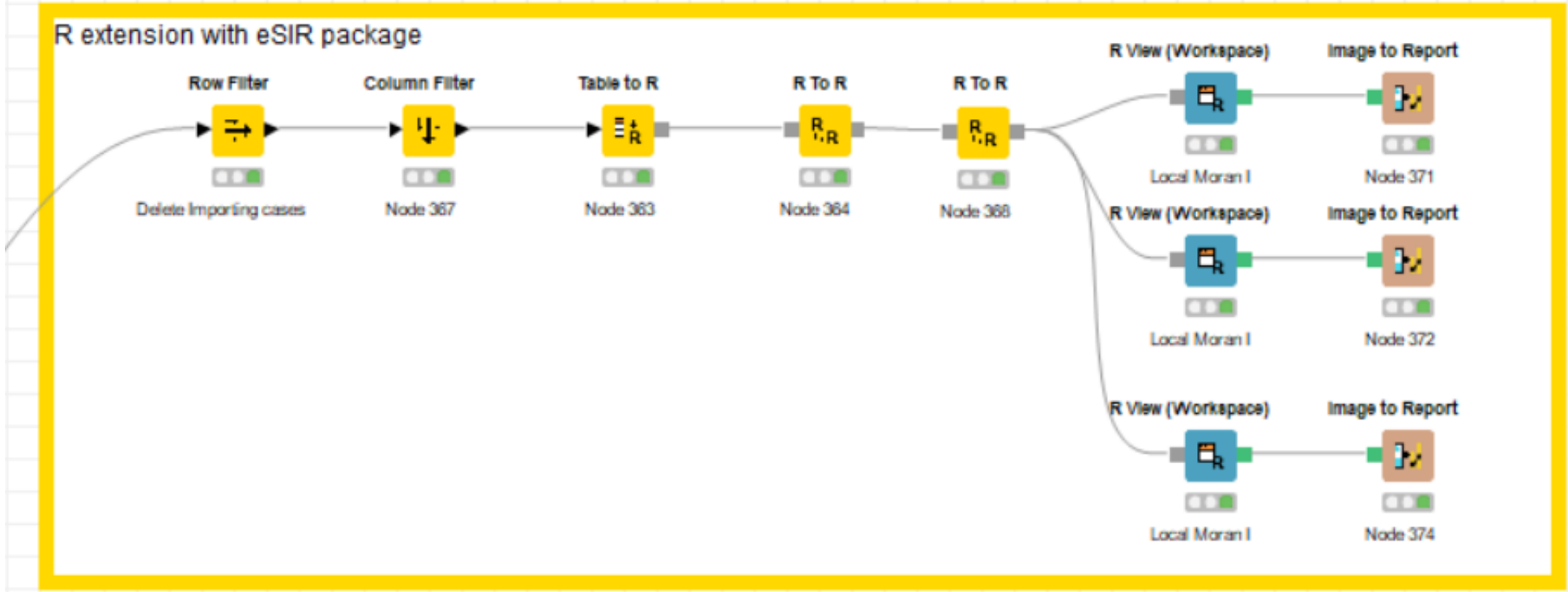


WORKFLOW FOR DATA ANALYSIS

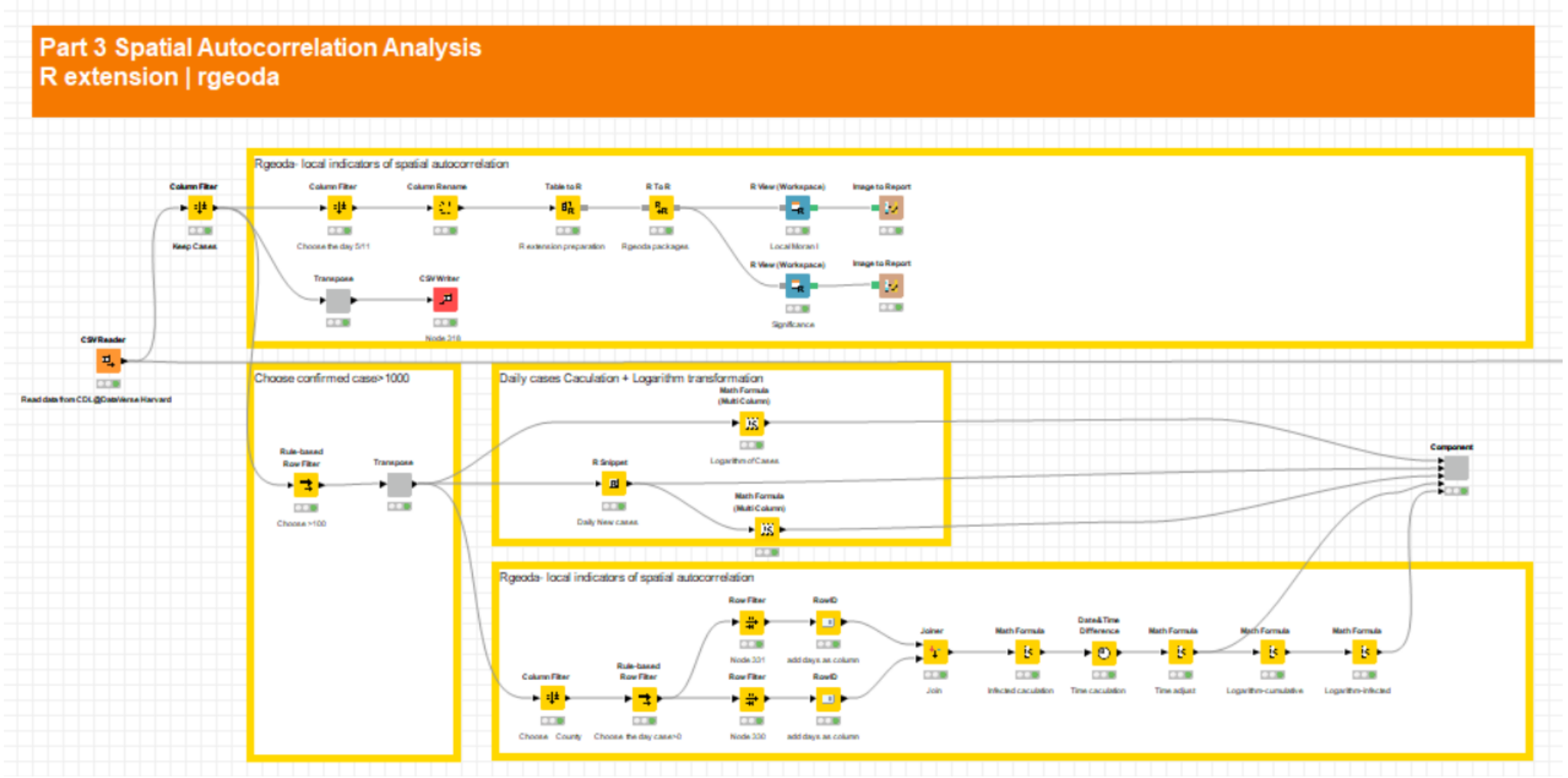


WORKFLOW FOR DATA ANALYSIS

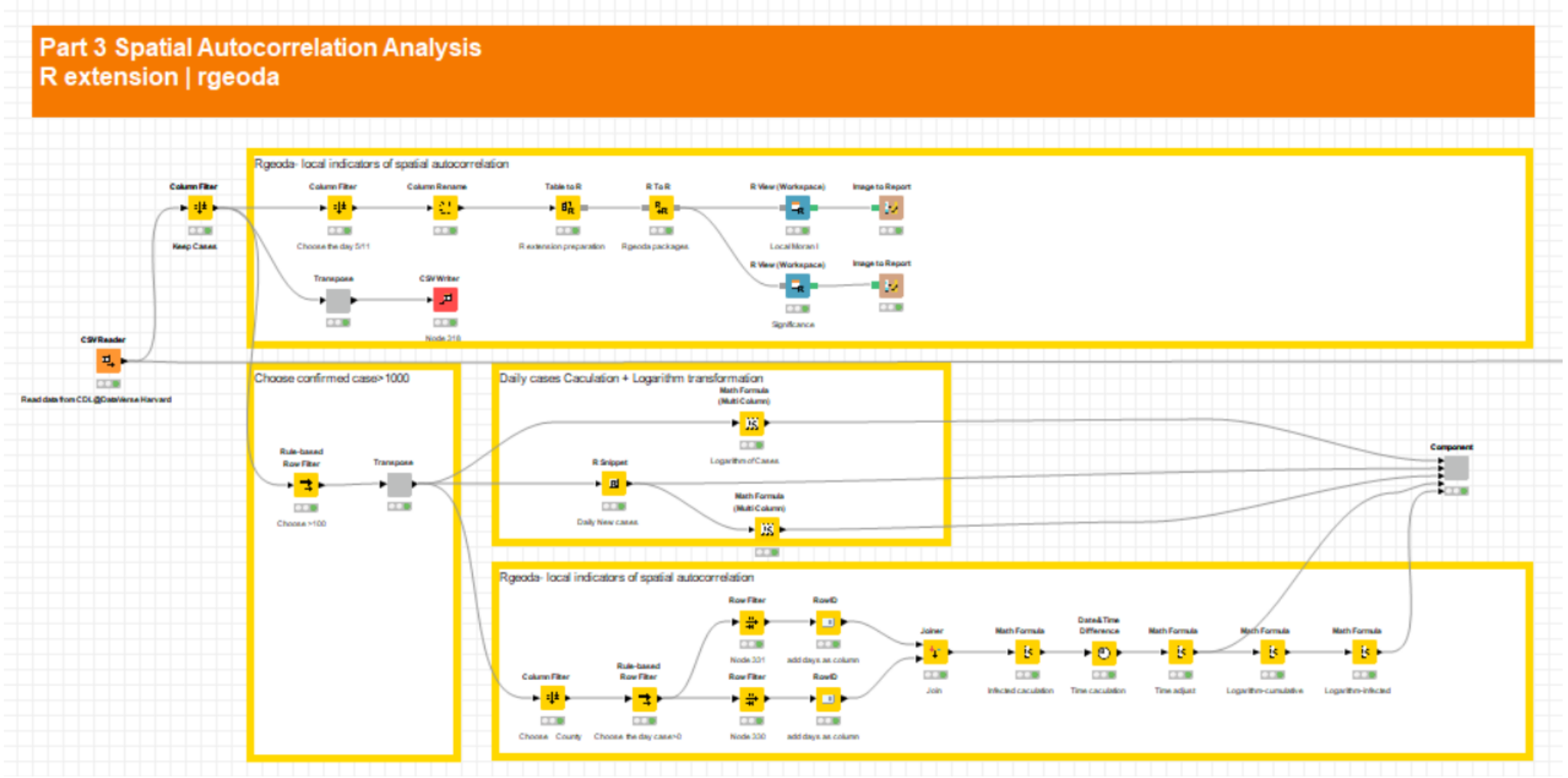
Part 2 COVID19 SIR model
eSIR package



WORKFLOW FOR DATA ANALYSIS

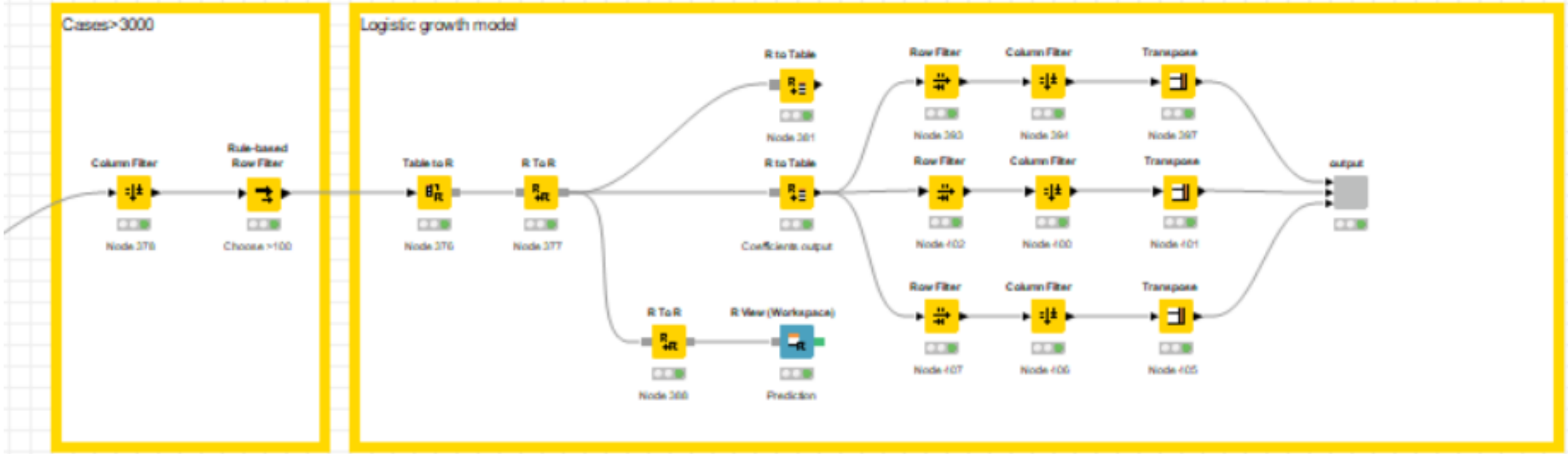


WORKFLOW FOR DATA ANALYSIS



WORKFLOW FOR DATA ANALYSIS

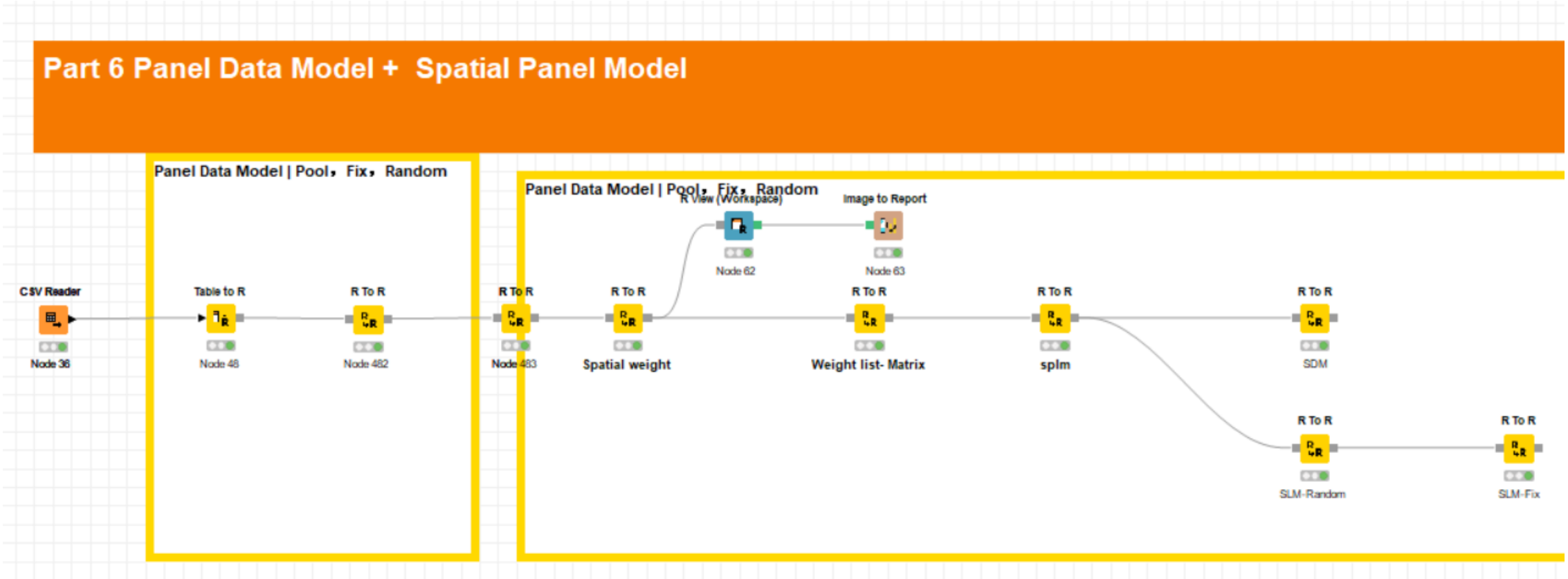
Part 4 COVID19 Simple SIR model
Logistic growth model+ SIR model



WORKFLOW FOR DATA ANALYSIS

WORKFLOW FOR DATA ANALYSIS

WORKFLOW FOR DATA ANALYSIS



Summary and Discussions

Spatial effect detected

The equation of Time varying transmission rate

Beta= $beta_0$^t

Spatial SIR need more investigation

More workflow could be developed