ASSIGNMENT 4: EXPLORATORY MODELING ANALYSIS

Variables of interest

- Main uncertain factors in this model: Birth rate, Death rate, Annual food requirement per person per year, Rain fall mm/year, Cross sectional area aquifer, Porosity, Average water used per person, Yield or irrigated crops, Rain lost, Irrigation water need per meter square, Average water use per person in industry per year, Average people per house, Average area per house, Average river flow AND Time of business demolishing.
- The range of each uncertain factor: This depends on the source that provides data. Most of those data are taken from governmental or organizations of United Nations to guarantee the reliability. First, in order to make a wide range of testing, I include the data from developing and developed countries, and used the maximum and minimum data from a findable source. The second reason of doing so is because of undefined location of current project. The phenomenon of salt intrusion in coastal area happens in everywhere, American eastern bank or Can Tho delta in Vietnam. So a wide range of testing would cover most of the possible cases.
- The statistical distribution of uncertain factor: Uniform distribution because this is the most reasonable distribution to use. Of course the distribution could be another type (normal distribution, for example), which depends on the data source. However, to simplify, in this case, uniform distribution is chosen so all values of factor range have equal chance to contribute to the calculation.
- The sampling technique chosen is Latin Hypercube because with a set 10,000 iterations, the sample withdrawn from data range will be represented more accurately compared to Monte Carlo sampling method.

List of Uncertainty Factor that have based on reliable sources

#	Uncertainty Factor	Value	Description
1	Birth Rate	• Initial: 4-6%/year • Range of testing: 0,685 – 4,76 %/year • Uniform distribution	According to The World Factbook[1], in 2012, the highest birth rate is 47,60 /1000 population (which means 4,76%/year) in Niger, and lowest birth rate is 6,85/1000 population (which mean 0,685 %/year) in Monaco (not including the negative birth rate).
2	Death Rate	• Initial: 5,5%/year • Range of testing: 0,155 – 1.723 %/year • Uniform distribution	According to The World Factbook[2], in 2012, the highest death rate is 17,23/1000 population (which means 1,723%/year) in South Africa, and lowest death rate is 1,55/1000 population (which mean 0,155 %/year) in Qatar (not including the negative death rate). Compared to Birth Rate, the Death Rate has a higher uncertainty level because there are more causes to death (natural death, age distribution, disease, war, etc…)

[1] https://www.cia.gov/library/publications/the-world-factbook/rankorder/2054rank.html
[2] https://www.cia.gov/library/publications/the-world-factbook/rankorder/2066rank.html

3	Annual food requirement per person per year	• Initial: 80 kg/person/year • Range of testing: 158 – 173 kg/person/year. • Uniform distribution	Let's assume that this is about cereal food. According to FAO (Food and Agriculture Organisation of United Nations)[3], the average cereal food consumption in 2015 in developing countries is 173 kg/person/year, in industrial countries is 158 kg/person/year.
4	Rain fall mm/year	• Initial: 150 mm/year • Range of testing: 150 – 1800 mm/year • Uniform distribution	According to National Rainfall Index (NRI)[4], the rainfall in over the world is categorised in different values from less than 300 till more than 1800 mm/year. It is able to collect this data from national climatic data centre with a relative high degree of accuracy. However, like other parameters, applying one value of climate to a large area could cause a big error. Moreover, in case of lacking data (some countries or regions do not have it), an estimation based on the available data of other countries or regions could be possible, though it takes high risks of error.
5	Cross sectional area aquifer	• Initial: of 3,65 x 108 m2 • Range of testing: ± 10%. • Uniform distribution	Because of the big magnitude of this factor, and of difficult in measuring it, this data should be considered as high or high uncertainty. Moreover, the project location has not been provided, so we do the test with range of testing of ± 10%.
6	Porosity	• Initial: 5% • Range of testing: 5-50% • Uniform distribution	Though this data does not have a big magnitude of value, it's still doubted about its accuracy. Is it possible to measure the porosity of one large area? According to some documents, this data could range widely from 5% - 50%.
7	Average water use per person per year	• Initial: 40 m3/person/year • Range of testing: 11 – 138 m3/person/year • Uniform distribution	The USGS Water Science School says[5] that in average an American consumes 80 – 100 US gallons of water at home every day. So maximum they use 138 m3/person/year. The lower bounder of the testing range (11 m3/person/year) is taken according to the data from the website http://chartsbin.com/view/1455 though it is not compatible to the above data.
8	Yield of irrigated crops	• Initial data = 1600 kg/m2/year • Range of testing = 0,2 - 1,2 kg/m2/year • Uniform distribution	Initial data = 1600 kg/m2/year Range of testing = 0,2 - 1,2 kg/m2/year According to the statistical data of cereal yield (kg per hectare) by World Bank in different regions or groups of countries per one session in year 2012: the minimum and maximum average yields respectively are 967 and 6341 kg/hectare/session (1 hectare = 10.000 m2)[6]. These figures do not

[3] http://www.fao.org/docrep/005/y4252e/y4252e04b.htm
[4] http://chartsbin.com/view/1425
[5] http://ga.water.usgs.gov/edu/qa-home-percapita.html
[6] http://data.worldbank.org/indicator/AG.YLD.CREL.KG

			mention anything related to irrigated or non-irrigated crops. Let's assume that one year there are 2 sessions of cultivation, the average minimum and maximum yields are around 0,2 and 1,2 kg/m2/year. So there is a crucial question to the validity of the initial values of crops productions.

The testing range

In this table, another 7 parameters (or uncertainty factors) are added. Due to the fact that no good data were found, the range of those parameters is estimated,

#	Uncertainty Factor	Initial Value	Range of testing
1	**Birth Rate**	4-6%/year	0.6852 – 4.7598 %
2	**Death Rate**	5.5%/year	0.1551 – 1. 7229 %
3	**Annual food requirement per person per year**	80 kg/person/year	158 – 173 kg/person/year. 100 – 200 (final testing range)
4	**Rain fall mm/year**	150 mm/year	150 – 1800 mm/year
5	**Cross sectional area aquifer**	3,65 x 108 m2	328504000 - 401496000
6	**Porosity**	5%	0.05002 - 0.49998
7	**Average water use per person per year**	40 m3/person/year	11006 - 137994
8	**Yield of irrigated crops**	1600 kg/m2/year	2.005 - 99.995
9	**Rain lost**	50%	60 – 80%
10	**Irrigation water need per meter square**	700 litters / meter square /year	350.0 – 1000.0

11	**Average water use per person in industry per year**	245 litre/person/year	245 - 2450
12	**Average people per house**	5	2 – 8 people/house
13	**Average area per house**	285 m2/house	150 – 400 m2/house
14	**Average river flow**	2×10^6 m^3/yr	2.100e+07 - 2.000e+10
15	**Time of business demolishing**	10 years	5.001 - 19.999

Initial settings

- The simulation number is chosen at 10.000 times with a uniform distribution of all 14 parameters (one among 15 parameters was omitted accidently).
- The noise seed is 1234 (default value)
- The sampling method is chosen as Latin Hypercube and multivariate.
- For each simulation, data of targeted variable "Crops produced per person per year) and 14 parameters are recorded (at 51 moments, year 0 to year 50).
- The data from Vensim is exported to R, converted into a data frame of 510,000 objects of 17 variables (Time [1:51], Simulation [1:10,000], targeted variable of "*Crops produced per person per year*", and the 14 parameters).

Data analysis

- **First**, the distribution of each parameter is checked by the histogram of values of parameters. To make every plot has the same scale, the data of parameters is exported to a separated set and then is normalized. For examples, the followings are 4 among 14 histogram plots. The rests have same distribution. Though it's not perfect uniform, the graphs prove that most value in each parameter range has the equal chance to contribute to the calculation.

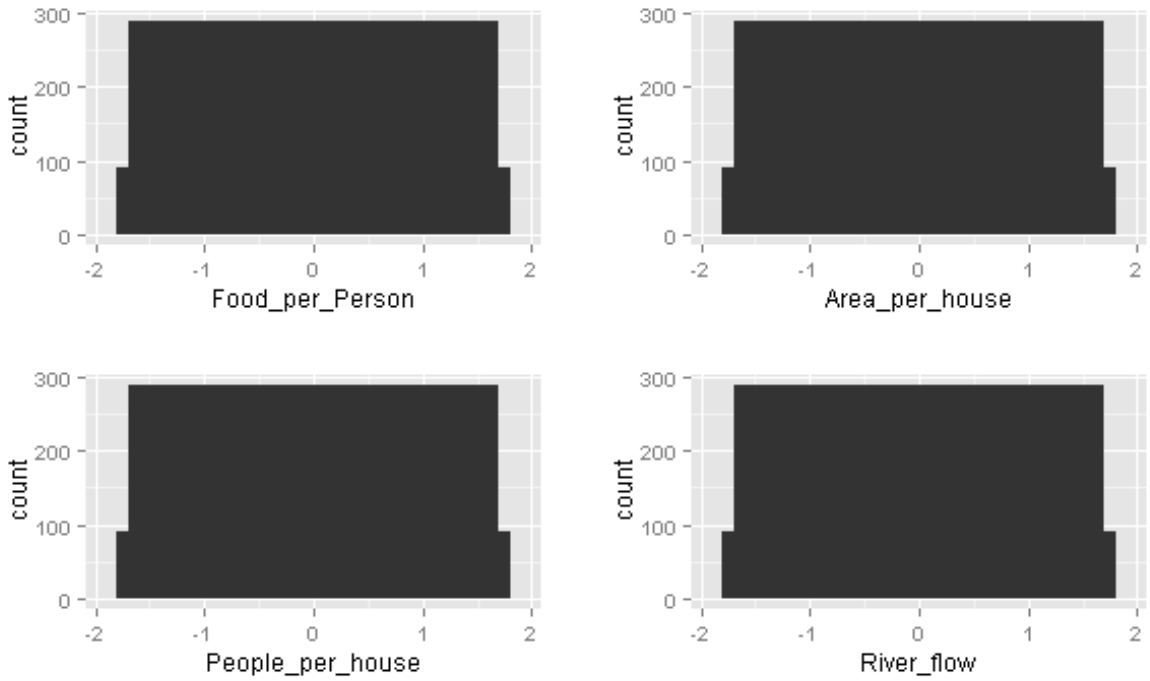

Figure 1: The distribution of 4 parameters (selected randomly among 14 parameters)

- **Second**, plot of "Crops produced per person per year" vs. "Time" is draw. The following plots consist of several cases: (1) combinations of plots all 10,000 simulations (2) combination of plots of first 4,000 simulations (3) combinations of plots of first 100 simulations.

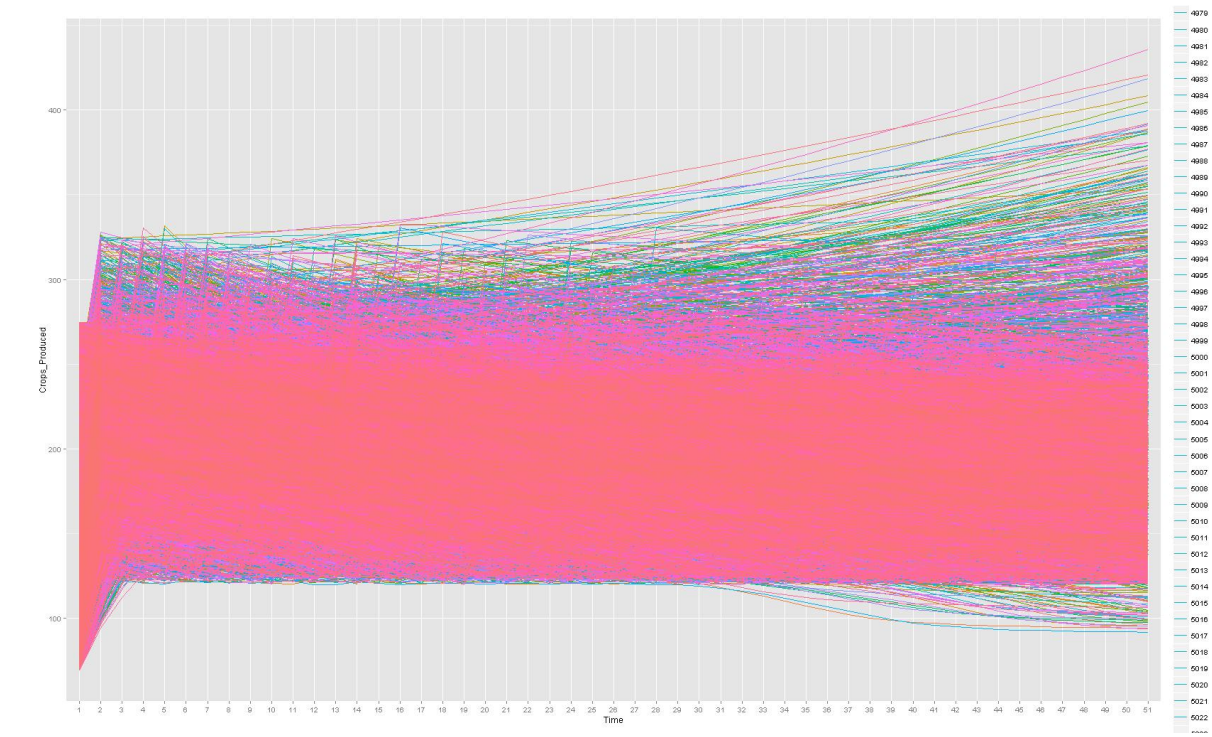

Figure 2: Combinations of plots of 10,000 simulations

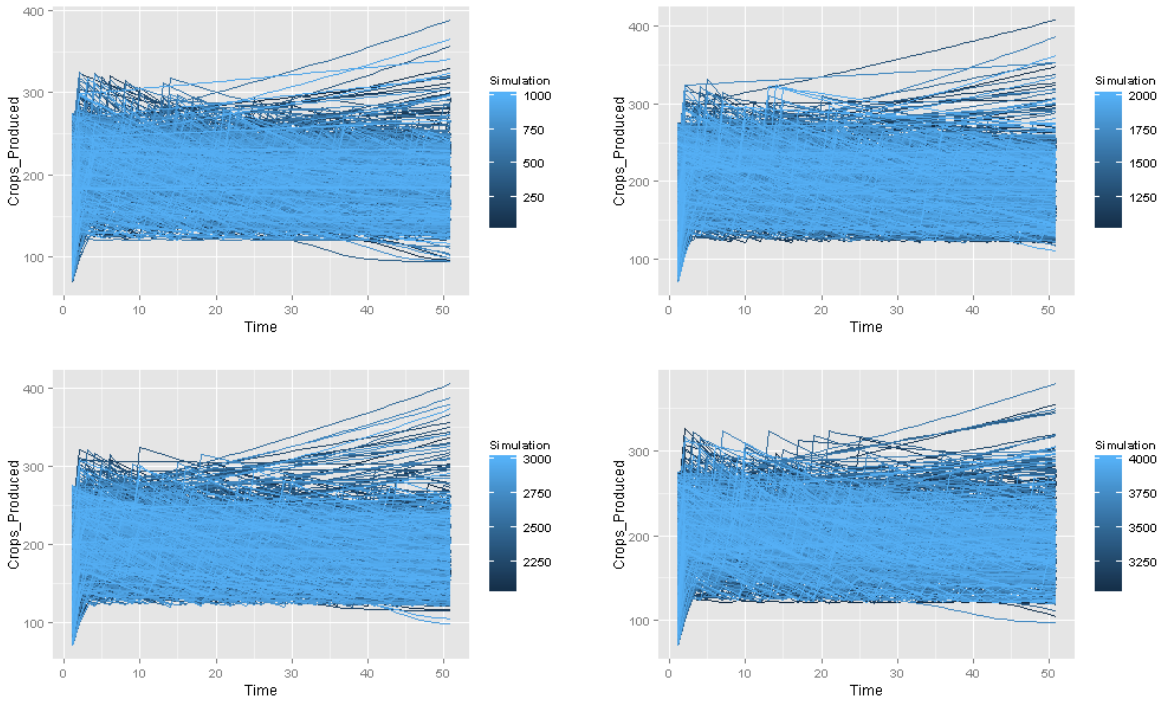

Figure 3: Combinations of plots of first 4,000 simulations

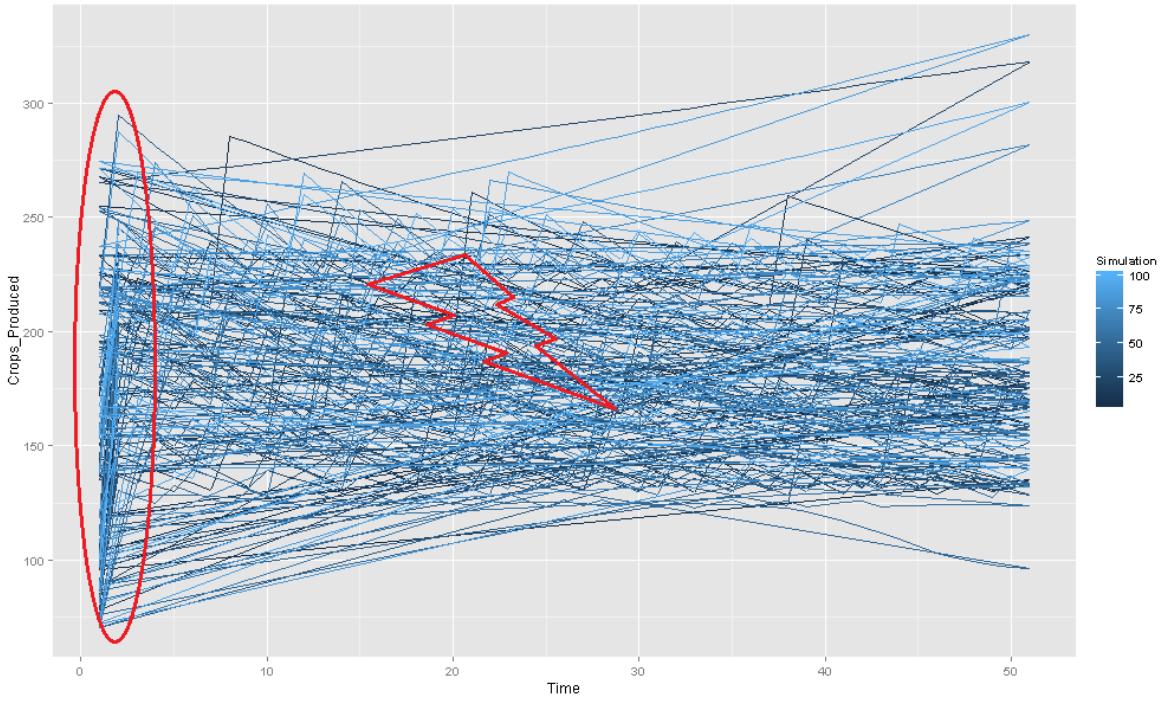

Figure 4: Combination of plots of first 100 runs

In general,

(1) Figure 2 (10,000 plots) shows that at the same time, different simulation corresponds to different values of variable "*Crops produced per person per year*". Some starts at higher values, some ends at lower values than the others. The densest part (with the pink colour) shows the most frequent happen of scenarios. However, there are still many deviations from this pink area.

(2) Figure 3 (4,000 plots) has the same message as the figure 3

(3) Figure 4 shows that in the whole timeline of 50 years, some simulations are flatter than others, some are very fluctuating.

Those preliminary observations lead to 2 broad questions:

(1) **Question 1**: Why at the start and at the end, some values of "Crops produced per person per year" are very high, the others are very low?

(2) **Question 2**: Why are some lines flatter or much less fluctuating than others?

To answer those questions, the main direction is to use Classification and Regression Tree (CART) technique as a tool of data mining. Others techniques are also applicable, such as AID and its variations, CHAID, etc. However, in this report, CART is chosen because (1) it is supported largely in R (2) Each method has its own advantages and disadvantages. In a scope of small project, one method is enough to demonstrate the picture.

Question 1

- "Why at the start and at the end, some values of "*Crops produced per person per year*" are very high, the others are very low?" In this report, we only conduct the measure and analysis at year 1. The same process can be applied at other moments, such as year 50.
- Data of year 1 is extracted into a subset, containing values of targeted variable *"Crops produced per person per year"*[7] at year 1 and 14 parameters for all 10,000 simulations.
- The histogram of targeted variable (figure 5) is plotted to illustrate the first view of distribution of its values. The plot shows that at year 1, the values of Crops_produced are partially normally distributed. 50% of cases have Crops_produced higher than 200 kg/person/year. Around 75% of cases have Crops_produced higher than 150 kg/person/year. Less than 20% cases have Crops_produced lower than 150 kg/person/year.

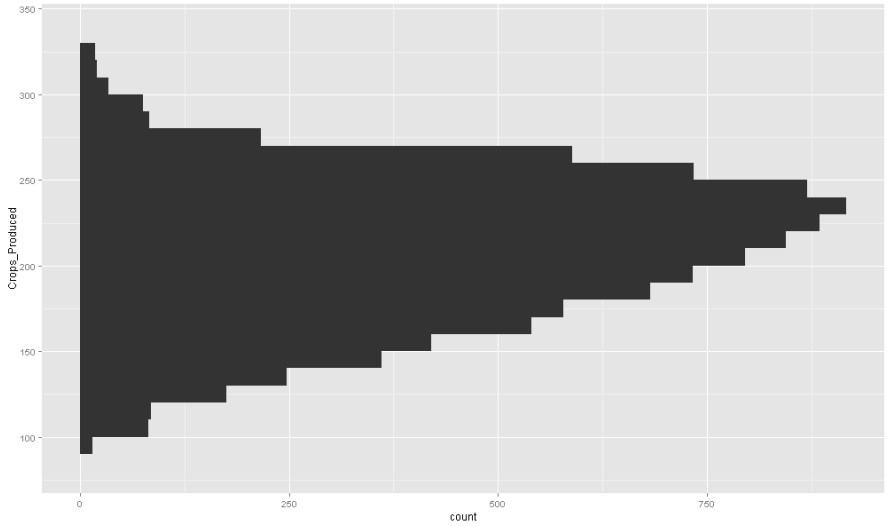

Figure 5: Histogram plot of Crops produced per person per year at year 1 for 10,000 simulations

- CART technique is applied to analysis the influence of 14 parameters on the distribution of Crops_produced at year 1. Because level of measurement of Crops_produced is ratio, "anova" method is used in CART technique settings. The predictors are 14 parameters.
- The result of CART analysis is illustrated by figure 6, 7 and table :

  (1) Only two parameters are significantly counted into the classification process. **(Yield of irrigated crops** (kg/ m2/ year) ~ crops_yi and **Annual food requirement per person per year** (kg/ person/ year) ~ Food_per)

Average Crops Produced Per Person Per Year	Number	Percentage
123	244	2
150	907	9
179	298	3
179	1350	14
123	213	2
164	247	2
209	1369	14
236	1506	15
214	1502	15
251	1412	14
262	952	10

Table 1: Categories of Crops Produced Per Person Per Year (Observations = 10,000)

[7] From now, it is written in short as "Crops_produced"

(2) More than 60% of 10,000 observations have average value of Crops_produced from around 200 to 260 kg/person/year. These results come from the branches that the Food Requirement Per Person Per Year ≥ 146 kg/person/year and Yield of irrigated crops > 62 kg/m2/year.

(3) Less than 40% of 10,000 observations have average value of Crops_produced from 123 to 164 kg/person/year. These results come from the branches that Yield of irrigated crops < 62 kg/m2/year and Food Requirement Per Person Per Year is less than 146 kg/person/year.

(4) In term of policy implication, the results provided a suggestion of an optimal combination of range of "Yield of irrigated crops" and "Food Requirement Per Person Per Year" to create a high "Crops produced per person per year. The "Yield of irrigated crops" could be improved quickly by new cultivation techniques.

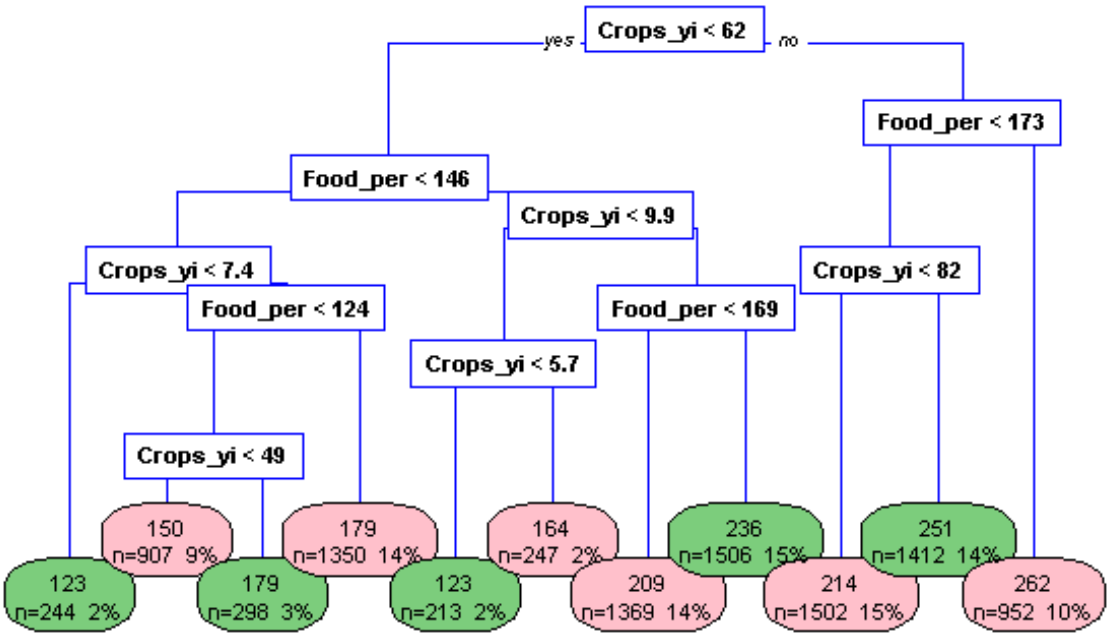

Figure 6: CART tree for question 1

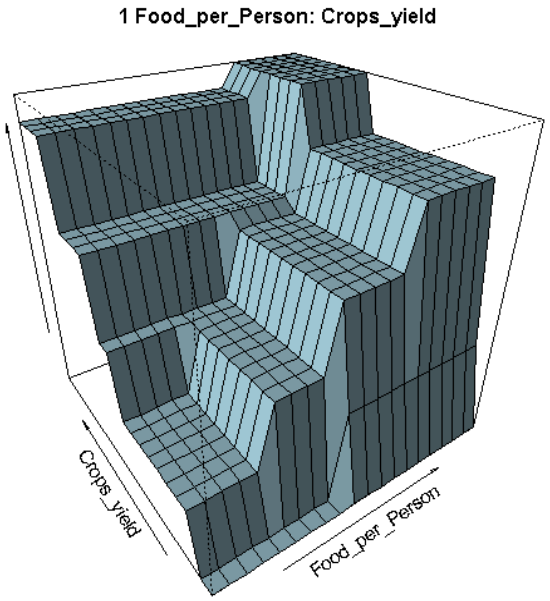

Figure 7: CART result for question 1

Question 2

- *Why are some lines flatter or much less fluctuating than others*?
- To answer this question, we analysis the time series values of each line, regarding to the derivative of whole line. At deeper level, the lines should be inspected why do they increase or decrease dramatically at some moment, that make the system really unstable. However, in this report, due to the limitation of capability, only derivative of whole line will be considered.
- The derivative of Crops_produced for a certain time step is calculated by comparing its current value to values of previous time step. Only absolute values of derivatives are used. Then the derivatives for whole time series (50 years ~ 50 values) are summed into one single value – ***total absolute derivative***. There are 10,000 simulations, so there are 10,000 "total absolute derivatives".
- The following figure 8 provides a first look of how the values of 10,000 "total absolute derivatives" distribute. There is approximate 50% of cases that total absolute derivate is higher than 200. When the total absolute derivate is above 200, the line is considered to be fluctuating; above 400 they are very unstable. There is around 25% of cases that total absolute derivate is lower than 100 which means the corresponding lines are more or less stable (or flat).

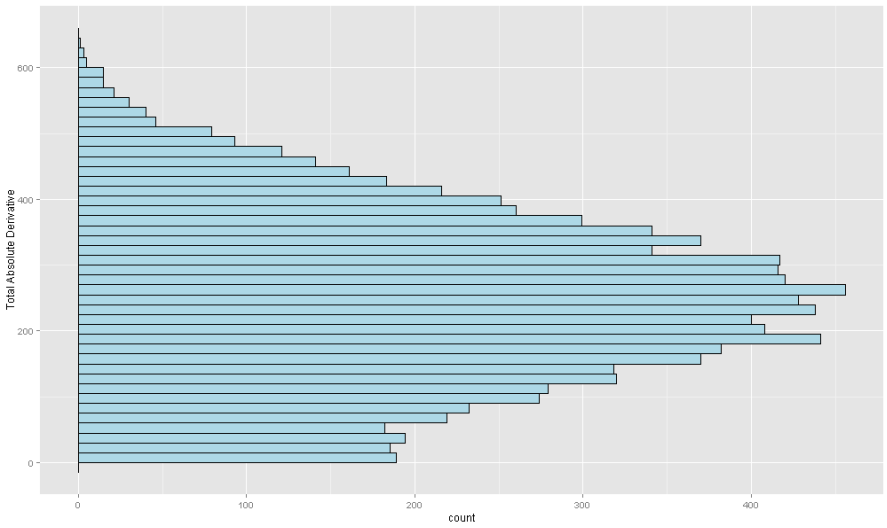

Figure 8: Histogram of "Total Absolute Derivative" for 10,000 simulations

- CART technique is applied to analysis the influence of 14 parameters on the distribution of "Total Absolute Derivative". Because level of measurement of this variable is ratio, "anova" method is used in CART technique settings. The predictors are 14 parameters.
- The result of CART analysis for question 2 is illustrated by figure 9, 10 and table 2:

(1) **Five** parameters are significantly counted into the classification process (Yield of irrigated crops, Annual food requirement per person per year, People per house, Birth rate and Death rate)

(2) The distribution of categories of Total Absolute Derivative is more fragmented than the case of Crops_produced in question 1. Perhaps this fragmented situation is resulted from the number of influencing parameters.

(3) Approximate 16% of cases have high values of total derivatives (which mean the lines are very fluctuating). This category corresponds to the branches where there are more than 3 or 4 people living in the same house (People_per > 3.6), Birth rate > 2.7%, Crops yield in irrigated land > 23 kg/m2/year and Food required per person per year > 146.

Average Total Absoluate Derivative	Number	Percentage
81	1322	13
140	559	6
146	924	9
180	601	6
214	761	8
248	766	8
252	324	3
297	1590	16
342	571	6
357	370	4
422	1564	16

Table 2 Categories of Total Absolute Derivative (Observations = 10,000)

(4) Approximate 13% of cases where lines are almost flat. This category corresponds to the branches when people require less food (less than 158 kg/person/year), high Death Rate (higher than 0.7% / year), low Birth rate (less than 2.1%).

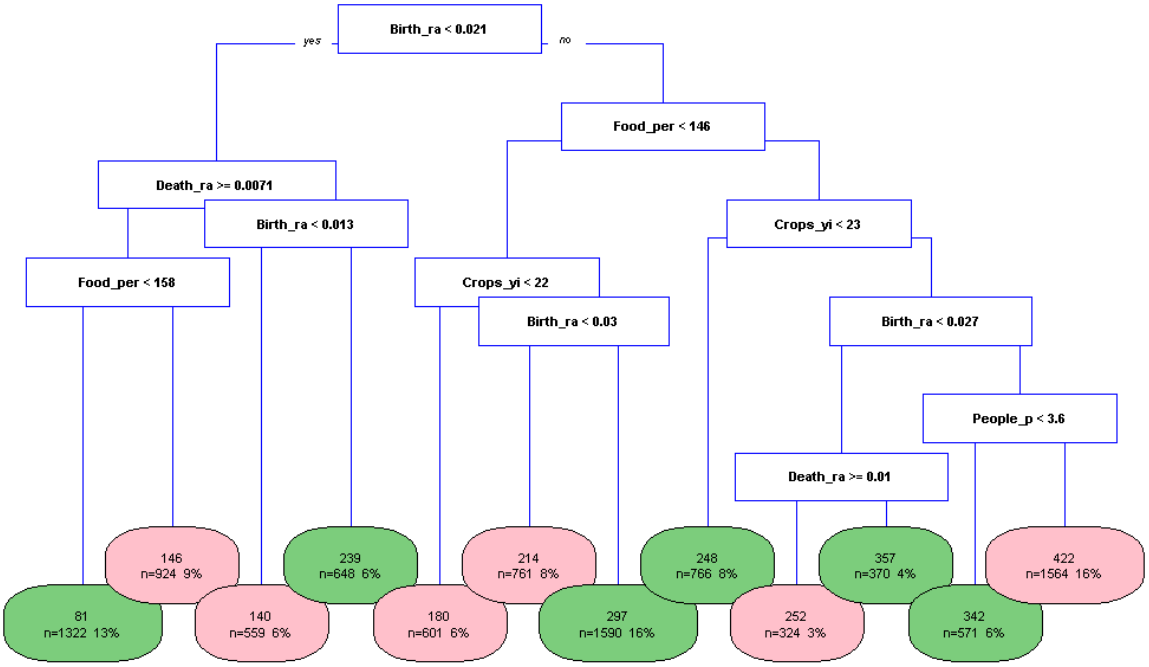

Figure 9: CART tree for question 2

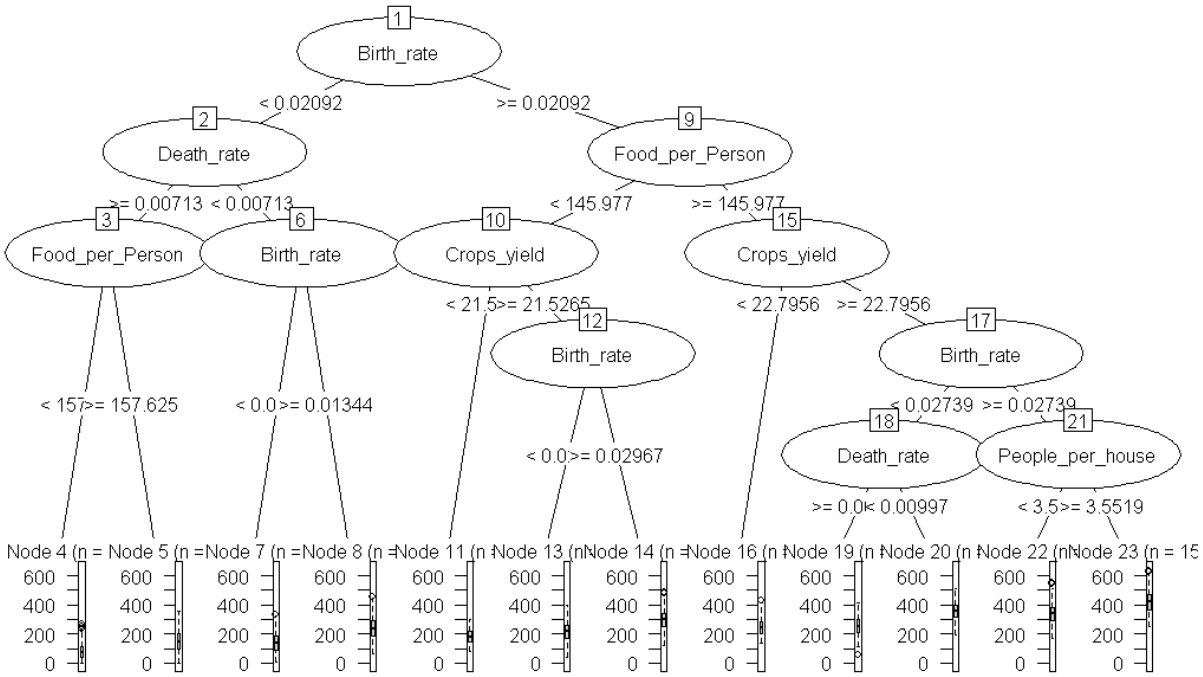

Figure 10: CART tree 2 for question 2

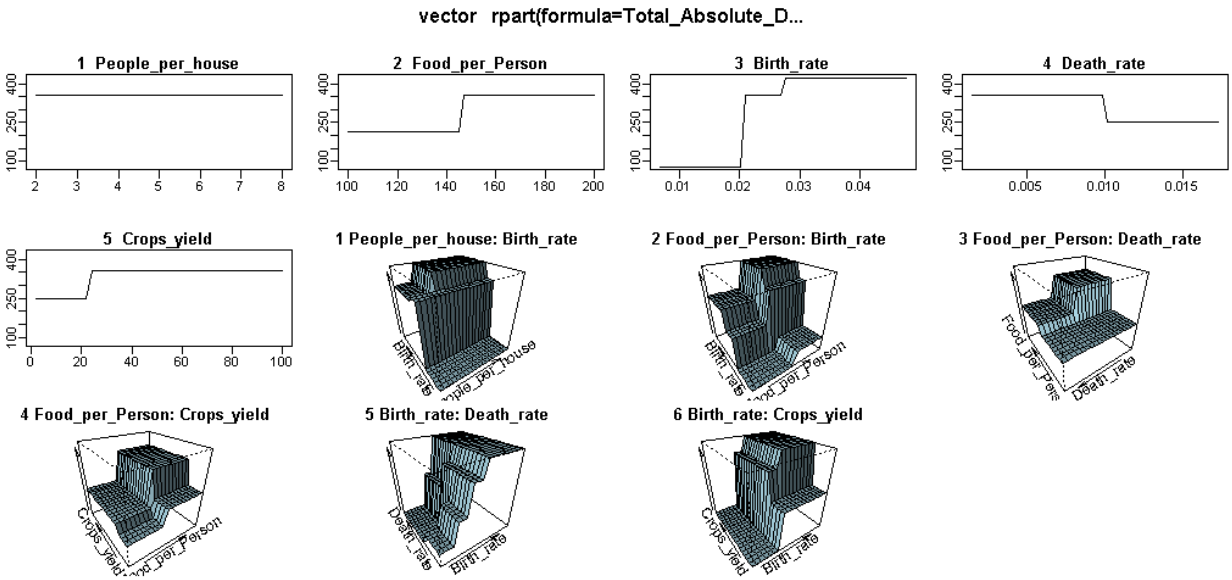

Figure 11: CART result for question 2

Conclusion

- Different from the sensitivity analysis where only few parameters were included in the experiment with a limited range of testing, EMA really explored and broadened the experiment into new scope. In this case, 10,000 simulations of model give us a dense picture with different patterns of variable "Crops_Produced" behaviors. By looking into those different patterns, new questions are more easily risen to investigate why there are more lines locating in the pink area, the others scatter sparsely below or beyond this area; why some are stable during the whole time-series why others keep fluctuated, changing direction strongly, etc.
- To answer the above questions, relations between variable "Crops_produced" and 15 uncertainty factors (parameters) must be revealed. However, because of big number of uncertainty factors and confusing (non)liner relation, regression tree (CART) is very useful to partition the space of "Crops_produced" and uncertainty factors into small regions. The final models that consist of limited factors involved (Crops Yield and Food Required Per Person for question 1; 5 parameters of Crops Yield, Food Required, Birth Rate, Death Rate and Number of People per House for question 2) help to:
  - quickly understand the results,
  - classify which parameters have important influences in overall, which behaviors are decisive, which new behaviors emerge and why?
  - explore new range of parameters that is not
  - really increase the perception of modeler from visualization perspective.
- Therefore, with regard to policy implication, modeler knows which parameters to focus in exploring the model (5 parameters in this case), in which behavior or scenarios they will involve, for which different ranges of 5 parameters will lead to different categories of final result of "Crops_produced". In brief, a wise combination of range of 5 parameters will help to build a robust policy in increasing crops produced per person per year while keeping this value stable and controllable.
- However, the work of EMA is highly labor intensive. (1) The data created by EMA experiment is big, so an appropriate tool to handle big data (Matlab, R, etc.) is needed. (2) Moreover, the EMA method also requires the skill of data mining or data analysis techniques (like CART). This technique is not relevant to System Dynamic traditional knowledge. Those 2 tool and techniques are crucial important during the process. Most of time is to handle, manipulate the data, highlight the focused area and apply CART technique in those areas.
- All of found-out results (for example range of Birth Rate, Death Rate and Food Required per person to stabilize the Crops produced) should be tested again, to check whether the policies really make change.