


Dancing with Data



Behind Hans Rosling's Graphs

Bradd Libby
May 2011

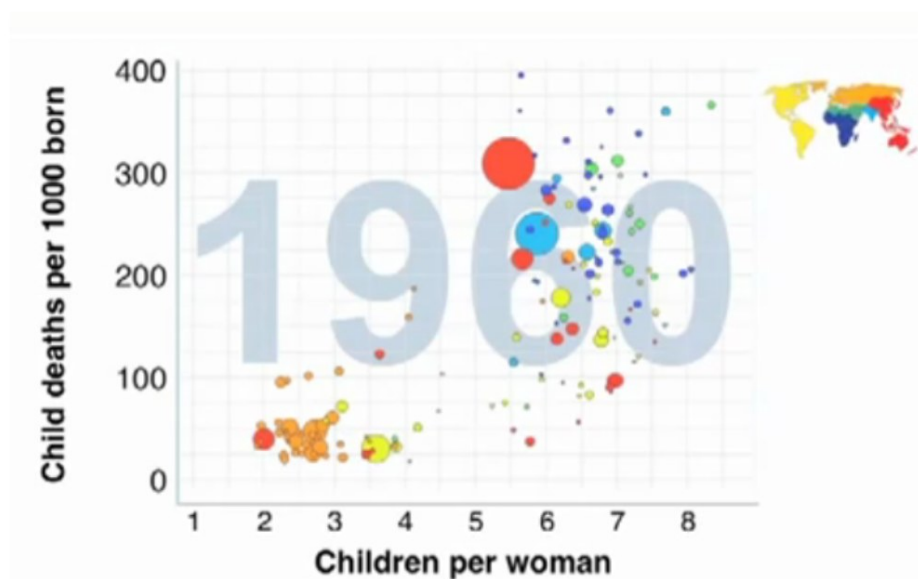


Let Us Now Praise Famous Men

I'm going to start and end this with only kind words of Dr. Hans Rosling. If you haven't seen his presentations from TED or on YouTube or the BBC, I highly recommend checking them out. There's always a place for popularizers of science and math, like Carl Sagan with astronomy and James Burke with technology, and Dr. Rosling is as good as any, though his field is global public health.

In several of his presentations, he has shown many animated bubble charts. For example in ['Good News of the Decade'](#) he focused on 'Child Mortality' vs. 'Children per Woman'. Each point on the graph represents one country during one year. By showing a series of plots, one for each year, we can see how different countries change their condition over time.

Here's a screenshot showing that graph for the year 1960:

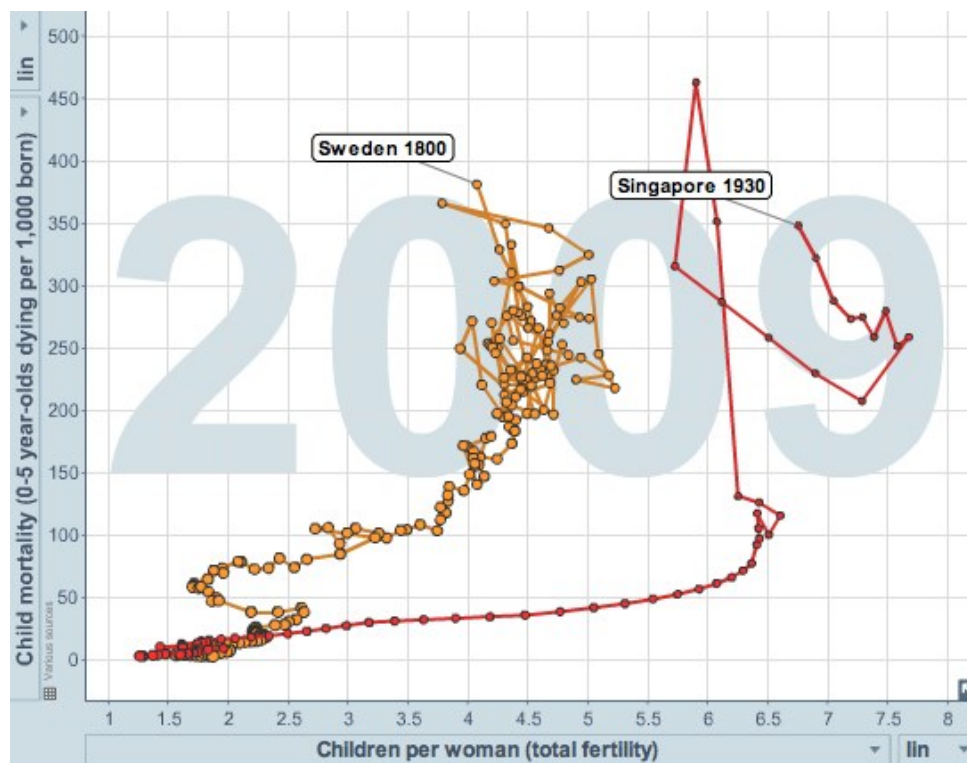


As late as about 1970 there were two distinct groups of countries: 'Developing Nations', which tended to have a larger number of children per woman (5 to 7) and higher child mortality rates (upwards of 50%), and 'Developed Nations', which tended to have a smaller number of children per women (1 to 3) and lower child mortality rates (under 10%).

However, in the past 30 years or so (with the notable exception of much of sub-Saharan Africa), this distinction has largely evaporated. Today, virtually all countries average 3 or fewer children per woman and have a mortality rate that is less than 10%.

Dancing with Data

Dr. Rosling compares Sweden (his home country) to Singapore, which since 2001 has had a child mortality rate equal to or lower than Sweden's and which averages even fewer children per woman. (Though, Dr. Rosling notes that, despite its better standing in terms of child mortality, a recent UN report still classifies Singapore as a 'developing nation'.)



If Dr. Rosling was less competent as a teacher, he probably would have just written a paper about this for an academic journal and then given a very boring presentation about it at a conference for scientists and world health officials.

But instead, he treats the development of these two countries more like a horse race: He sets the graph into motion, having each country leave a colored trail as the years tick by.

Because the projection screen is much larger than he is, Dr. Rosling needs to run around as the graph updates. As child mortality bounces around, he must jump up and down with it. And as the number of children per woman drops over time, so too must Dr. Rosling run from one side of the stage to the other. Hans Rosling doesn't just point to the data - he *dances* with it.

As each country's fortunes shift, Dr. Rosling excitedly notes the change in momentum. And when, in the last few years, Singapore "caught up" (as Dr. Rosling says) with Sweden, he breathlessly proclaims that like the announcer at the Kentucky Derby calling a photo finish.

Singapore's performance is remarkable, being in 1930 way above Sweden in both children per woman and in child mortality, but being lower in both metrics just one lifespan later. This seems to be the reason why Dr. Rosling feigns disgust when tearing the page from the UN report that still lists Singapore as a developing nation. "This is crap!", he gushes, as he throws the crumpled paper to the floor.

If Dr. Rosling were a younger, less-distinguished researcher, a tenure committee might conclude that he was not a 'serious' academic. A serious researcher would buy a laser pointer and talk in a monotone. But Dr. Rosling is serious about revealing the inner beauty of data and making it available for all to see. *Very serious.*

This, I think, is the reason why Dr. Rosling built [Gapminder](#), a software tool that generates animated bubble charts like the ones he uses in his presentations, collecting together data from many countries, for many years, on not just these two factors - 'Child Mortality' and 'Children per Woman' - but hundreds of others, covering health, wealth, the environment and human well-being. The history of world can dance before your eyes, free for anyone who has both a web browser and the motivation to explore.

Mine the Gap

As entertaining as his work is, though, there is something about presentations and software, soon after the dance ends, that leaves me with a feeling of dissatisfaction. The Gapminder is full of data on all sorts of dimensions - Gross Domestic Product, total population, average education levels, CO₂ emissions per person per year - and it's very easy to use the software to find correlations, such as life expectancy (in years) tending to rise among nations as per-capita income (in inflation-adjusted dollars per person per year) rises.

But in the end, the correlations we see on bubble plots are just that, *correlations*. No separation of cause from effect. No explanation deeper than: "This is what the data shows."

Look back at the graphs of 'Child Mortality' vs. 'Children per Woman' above, for both the

entire world in 1960 and for Sweden and Singapore throughout many decades. Here's one question I have: **Why do the relationships (in both graphs) have that scoop-like shape?**

For the first graph we're talking about the *entire world* here. There are Communist countries in there and Capitalist ones. Democracies and dictatorships. Those that have well-developed public health systems and those that do not. Countries near the Arctic and those near the Equator. Former Colonizers and former Colonies.

And yet they all seem to align themselves along the same curve.

The graphs for Sweden and Singapore each have essentially the same shape as each other too, even though I can't think of anything they have in common except that they both start with the letter 'S'. This seems interesting to me, but Dr. Rosling makes no effort to explain why Singapore caught Sweden or how to actually help Senegal, for example, catch Spain.

Do falling mortality rates prompt women to choose to have fewer children, or does having fewer children mean that women are better able to care for the ones they do have, making those children more likely to survive? Or, are there other factors, like rising per-capita income or urbanization that caused the observed change in both 'Child Mortality' and 'Children per Woman'? From Dr. Rosling's graphs we just can't tell.

From Correlation to Causation

And what could we do if we wanted to help make Sierra Leone (to name another 'S' country) more like Singapore? Dr. Rosling claims that the Institute for Health Metrics found that almost 50% of the observed fall in child mortality can be attributed to increases in female education. But what effect does an increase in female education have on the number of children per woman? Again, we just can't tell from Dr. Rosling's graph.

Instead, we need a model - one that gets us away from the observed relationships we see in bubble charts and towards the 'social mechanics', you might call it, that generated that observed behavior.

A Simple Model of 'The Good News of the Decade'

I'd like to propose what (I hope) is a simple mechanism that demonstrates some of the same behavior Dr. Rosling saw in his graphs. Suppose that there are two types of women: for lack of better expressions, let's call them 'More-Educated' and 'Less-Educated'. More-Educated women have 2 children (half of whom are female), one at age 30, one at 32. Further, 5% of their children die young. All of their surviving daughters become More-Educated.

In contrast, Less-Educated women have one child every other year from the ages of 20 through 32. (Thus, they give birth 7 times.) Before year 1900, 40% of these children die, but after year 2000, only 5% die. In between 1900 and 2000, the mortality rate drops each year and in this investigation we'll explore what happens if the rate of decrease is fast or slow. Before 1900, none of their daughters become More-Educated, but after year 2000, all of them do. Between 1900 and 2000, the fraction who become More-Educated changes. Again, we'll explore faster and slower rates of change.



More-Educated Women

- have one child at age 30 and another at age 32
- child mortality rate is 0.05
- all surviving daughters become More-Educated women

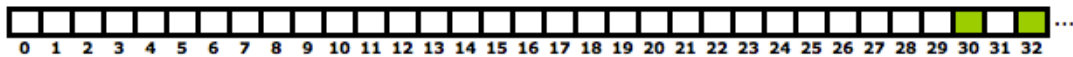


Less-Educated Women

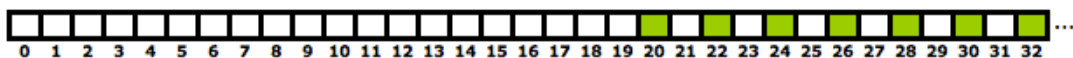
- have one child every other year from age 20 to age 32
- child mortality rate is 0.40 before 1900 and is 0.05 after year 2000
- some surviving daughters become More-Educated (the rest become Less-Educated)

I think of the population of females of each type to be like a conveyor belt. Each segment of the conveyor contains the number of females of a particular age. Each year, the number of women at each age advance to the next age. So, in the year 1950 if there are 1000 More-Educated women who are 26 years old, then in 1951 there will be 1000 More-Educated women who are 27 years old.

'More-Educated' females



'Less-Educated' females



Also each year, each More-Educated female who is 30 or 32 years old gives birth to 0.5 females. (Of these, 5% die immediately and the rest become More-Educated newborns.)

And, each Less-Educated female who is age 20, 22, 24, 26, 28, 30 or 32 also gives birth to 0.5 females. Of these, some percentage (the Mortality Fraction) die immediately. Of the rest, some percentage (the Education Fraction) become 0-year-old *More-Educated* females.

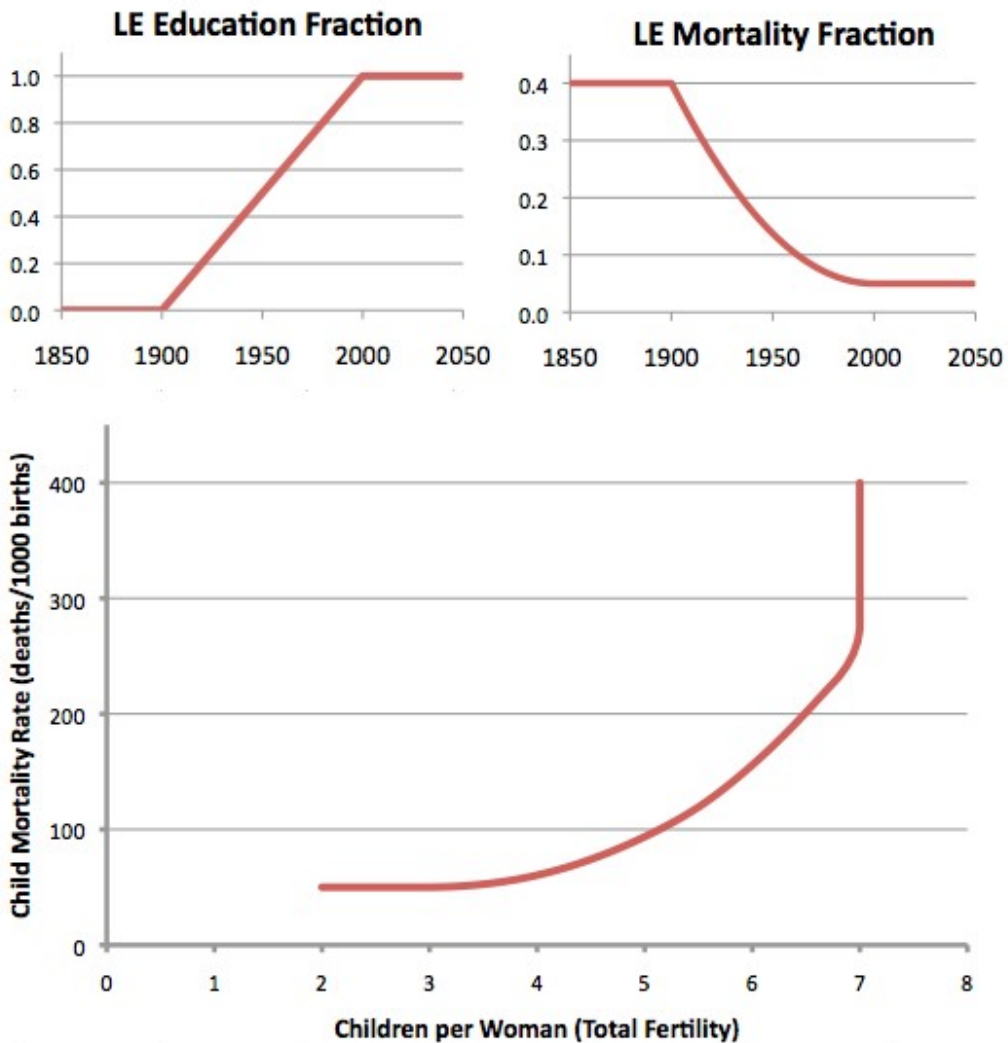
The **Mortality Fraction** is 0.40 before the year 1900 and is 0.05 after the year 2000. Between the years 1900 and 2000, the Mortality Fraction drops from 0.40 to 0.05, but the specific rate at which it drops can be changed to be faster or slower.

Likewise, the **Education Fraction** is 0.00 before the year 1900 (that is, all of the daughters of Less-Educated women also become Less-Educated women) and is 1.00 after the year 2000. Between the years 1900 and 2000, the Education Fraction rises from 0.00 to 1.00, but the specific rate at which it rises can be changed to be faster or slower.

The calculations start in the year 1800, when there are no More-Educated females of any age and there are 10 Less-Educated females at each age category. We advance the simulation 100 years before showing any results, in order to eliminate the effect of this simple way of initializing the population. The simulation is then run for 150 years, from 1900 through 2050.

Better Red than Dead

Imaging there's a country (call it 'Red Country') where the fraction of Less-Educated women whose daughters become More-Educated (that is, the Education Fraction) rises linearly from 1900 to 2000. Also, the fraction of Less-Educated womens' babies who die in infancy (the Mortality Fraction) drops from 0.40 to 0.05 according to $(1-t)^2$, where t is the fraction of the way the year is between 1900 and 2000. [If this description is confusing, the graphs below show the changes in Education Fraction and Mortality Fraction for Less-Educated women.]



If this is how the Education Fraction and the Mortality Fraction of Less-Educated womens' daughters change In this country, then the relationship between the Child Mortality Rate and the number of Children per Woman will look like scoop-shaped curve seen above.

Prior to 1900, there are no More-Educated women. The Child Mortality rate is 0.40 and the average woman has 7 children in her lifetime. Starting in 1900, health and welfare improvements for Less-Educated women cause the Child Mortality Rate to begin to fall. In say 1905 or 1910, there are now More-Educated females for the first time, but the oldest among them are still only 5 or 10 years old. So, among women in their child-bearing years, there are still only Less-Educated women, so the number of Children per Woman has not yet begun to fall.

Only when the first More-Educated females hit their 20's and keep getting educated when they would otherwise be having children (if they had instead been Less-Educated) does the number of Children per Woman begin to fall. As the 20th Century progresses, Red Country moves toward the end-state of having a Child Mortality Rate of 5% and 2 Children per Woman as all of the women in the population become More-Educated.

If all the countries of the world were like Red Country, but began their development before or after 1900, then they would all be arrayed somewhere along this curve and this might explain why Dr. Rosling's bubble plot for the world has every place from Sweden to South Africa falling in that scoop-shaped cloud.

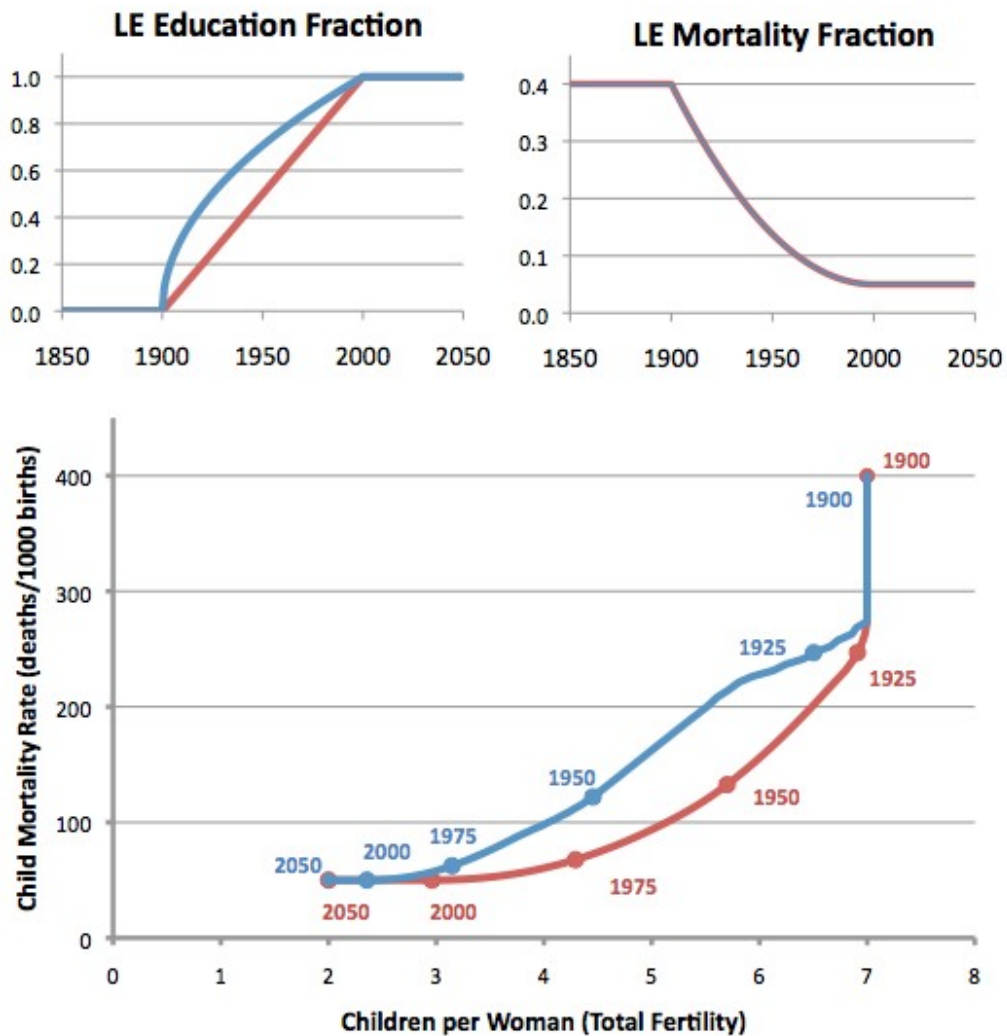
In a Blue Country

The education of young women has gotten enormous attention in the past years, not only due to people like Dr. Rosling, but also to people like Greg Mortensen (author of *Three Cups of Tea*), who has come under some scrutiny recently but whose motivations seem good-hearted: More-Educated women tend to have a lower child mortality rate, it is reasoned, so educating young girls should tend to decrease the child mortality rate. With my model, we can test this, to see how much child mortality should drop as education rates rise.

"We in public health," Dr. Rosling says in his Good News of the Decade presentation, "we hate child death. We want less and less and less of child deaths."

Imagine there's a second country, 'Blue Country', identical in every way to Red Country except that the Education Fraction rose faster from 1900 to 2000. I've used the same mathematical relationship as for the Mortality Fraction just flipped upside down. (Again, it might be easier to look at the graph.)

What effect does increasing the Education Fraction have on child mortality in Blue Country compared to Red?



We can see that, again, Blue Country sees its child mortality rate drop in the first years after 1900 and in the long run of course, it also approaches 5% with a number of children per woman of 2. But the path that Blue Country, the country that educates its girls more aggressively, takes is otherwise very different.

For any given level of Children per Women, the Child Mortality Rate in Blue Country is higher than in Red Country. **That is, for any given level of Children per Women, raising the fraction of the daughters of Less-Educated who become More-Educated tends to *increase* the Child Mortality Rate, not decrease it.**

I find this very counterintuitive and it was not easy for me to grasp the concept. Perhaps think of it like one of Dr. Rosling's horse races: There is a starting point in 1900 with a 40% child mortality and 7 children per women. And there is an ending point with 5% child mortality and 2 children per women.

The shortest distance between two points is a straight line and the path that Blue Country took is closer to a straight line than Red Country's path. Thus, Blue Country is able to get near the ending point faster than Red Country.

Actually, we can find the optimal path by having Education Fraction go immediately from 0 to 1 in the year 1900. In that case, the country follows the same path as Red and Blue for the 'lag period' of the first 20 years, but then the Child Mortality Rate stays constant at about 250 deaths per 1000 births as the number of Children per Woman drops from 7 to the replacement level of 2. That is, for virtually any number of children per woman, the 'best' child mortality rate is about 250, not 0.

If that idea is puzzling, keep in mind that in this 'best case' scenario, the country would only have a number of Children per Woman that is a value between 7 and 2 for a period of about 10 years. That is, even though the 'best' child mortality rate is about 250, we should only stay at that level for a very short period of time in order to race as quickly as possible from 7 children per woman to 2 children per woman.

I've marked some selected years on each graphs and you can clearly see that, even though Blue Country has a higher child mortality rate at any given number of children per woman, Blue clearly outpaces Red. In 1925, the two nations are very similar in both child mortality and children per women. By 1950, Blue is significantly lower children per woman, but child mortality is not much different. By 1975, though, Blue Country (the one that has been educating its girls more aggressively for 75 years), it at a position which it will take Red Country about another 25 years to achieve.

Let's Do It

Here's one thing that bothers me about Dr. Rosling's presentations: we can see what *did* happen with Sweden or Singapore, but not what *could* have happened nor what *should* have happened. Dr. Rosling can only give the 'Good News of the Decade' talk once before he has to find another graph to dance with. But with a simulation, we can adjust our assumptions and re-calculate the results to see the impact. (What if More-Educated women only had 1 child instead of 2?)

Here is something that bothers me about the Gapminder: it seems like a lot of work. If we want to know the fraction of women who are More-Educated, we have to make a new graph in the Gapminder. Someone else needs to have already collected that data. And if they haven't, we can't make the graph. But with my model, we can generate those graphs ourselves simply by dividing the number of More-Educated by the total population and then show the results. If we want to know the distribution of women in each age group, we simply output the numbers from the simulation, since they are already there.

With a robust model (even a simple one), we can also add other factors to see how they affect and are affected by factors that are already in the model. In his 'Good News of the Decade' presentation, Dr. Rosling showed a graph summarizing the UN Millennium Development Goals that looked very similar to this one:



The UN makes it sound like these are 8 separate things. They even use different colors to distinguish them. But promoting the education of girls (part of Goal 2) affects Goal 4, for

example. And one thing that isn't apparent from the Child Mortality vs. Children per Woman graphs above is that, in my example above, the country that promoted education, the Blue Country, has a total population that peaks at a level much lower than the Red Country's population. That is, by promoting the education of girls, we help address the population issue at the same time.

As I said, I would start and end this by praising Dr. Rosling. He's a smart man and he understands these interdependencies very well. He ends 'Good News of the Decade' by saying "we can't solve the climate crisis without stabilizing the population". To do this, he suggests, we need to get child mortality down, get access to family planning (to reduce the number of children per woman), and drive female education. "That is fully possible. Let's do it!" he exclaims.

But to do that, we need to do more than wave our hands in front of graphs. Animated bubble plots are a good start to see what has happened. But to see what can and should and will happen, I will always find making causal models like the one I've described to be much more interesting. And I hope you will too.

Appendix: Re-creating the 'Good News of the Decade' (GOND) Model

Create two arrays with 50 boxes each. One array represents Less-Educated females, the other represents More-Educated females. Each box represents one age category. Start at year 1800 with no More-Educated females and 10 Less-Educated females at each age.

Each year, the number of females in box N becomes the number from box N-1. (So, in 1801, the number of Less-Educated females who are 27-years-old becomes the same as the number who were 26-years-old in 1800.)

Each year, each More-Educated woman who is 30- or 32-years-old gives birth to a baby, half whom are girls. 5% of these babies die immediately. The other 95% of these girls become More-Educated females in box 0.

Each year, each Less-Educated woman who is age 20, 22, 24, 26, 28, 30 or 32 also gives birth and again half are female. In the year 1900 and before, 40% of the babies die immediately. In the year 2000 and after, 5% of the babies die immediately. In between 1900 and 2000, find the portion of time that has elapsed since 1900 – That is, take the current year, subtract 1900, and then divide by 100. So, in the year 1953, the number (t) is 0.53. The fraction of babies who die is $[0.05 + 0.35(1-t)^2]$.

Of the daughters of Less-Educated women who survive their birth, the fraction who become More-Educated females of age 0 is t^x (where x is 1 for Red Country and 0.5 for Blue Country). The remainder of the daughters become Less-Educated females like their mothers.

Calculate the average Child Mortality Rate by adding up all of the babies who are born but die immediately and dividing by the total number of births.

Calculate the number of Children per Woman (actually, the Total Fertility Rate) by summing up, for each age where women have children, the ratios of the number of babies born that year to women that age to the number of women that age.

If find anything interesting by making useful changes to anything above, please let me know by email at bradd_libby@yahoo.com