Defecting from Autocracy:

Evidence from North Korea and a Simulation Model[1]

Alexander Dukalskis, University College Dublin
Johan A. Elkink, University College Dublin
Hang Xiong, ETH Zurich

11 June 2017

Work in progress

Prior to the 1990s defection from North Korea was extremely rare. This changed with the famine that devastated North Korea during the mid-1990s and today over 30,000 North Koreans reside in South Korea while perhaps as many or more live in China. This paper proposes an agent-based simulation model to explore the dynamics of defection from North Korea. Methodologically, the challenge lies in the small proportion of the actual population that defects. Defection is a contagious process, whereby successful defections encourage further defections, but nevertheless remains at an extremely low rate. To model a realistic proportion of defection, the implementation of the simulation will have to allow for a large number of agents, whose behavior will have to be simulated. We model the entire North Korean population based on demographic and economic characteristics, as well as national economic dynamics. We investigate the direct and indirect impacts of the famine that occurred in the mid-1990s nationwide as a systematic shock, which substantially changed the landscape of defection from the country in its aftermath. Simulations are run for a large number of replications to evaluate the robustness and consistency of the simulation findings and are calibrated to the empirical record of defection from North Korea from 1992 to 2014. The simulation results suggest that the North Korean regime has been reducing its repression levels over time, but also that originally individual North Koreans were relatively optimistic about their chances of success and that based on observed defection attempts, they have updated their expectations to more accurately reflect the actual levels of repression. It suggests that citizens are aware of about half the attempts at defection – the remainder goes unnoticed – and that they base their assessment on defection attempts over the previous seven or so years.

[1] Paper to be presented at the International Conference of the International Studies Association in Hong Kong, 15-17 June 2017.

Introduction

The 1990s were a challenging time for the Democratic People's Republic of Korea (DPRK or North Korea). As the world watched European and Soviet communism collapse many observers anticipated the same would occur in North Korea. In addition to its allies and trade partners vanishing, Kim Il Sung, the country's leader of nearly 50 years, died in July of 1994 after which his significantly less charismatic and credentialed son, Kim Jong Il became the DPRK's top leader. South Korea, the North's arch-rival had democratized and become a prosperous country, thereby undermining the North's ability to plausibly claim that the South was poor and ruled by a repressive military dictatorship. Perhaps most consequentially, North Korea suffered a devastating famine in the mid-1990s in which up to 1,000,000 people died of starvation and starvation-related causes.

Despite these challenges to its rule, the autocratic government in Pyongyang still sits atop and permeates what is perhaps the most rigidly controlled and repressive society in the world (United Nations 2014). Surveillance is pervasive and despite the recent emergence of the illegal or quasi-legal 'second economy' the state retains formal control over virtually the entire economy (see Joo 2010). The education system is designed to produce pliant and loyal citizens and the public sphere is thoroughly dominated by the ideology of the state (see Hassig & Oh 2009; Dukalskis 2017). Labor camps imprison up to 200,000 individuals for a range of both ordinary and political offenses (Hawk 2012).

Under these conditions people who wish to change the status quo have few options. There is no electoral arena in which they organize to pressure their leaders and extra-systemic collective action that questions government rule is mortally dangerous. If they remain in North Korea they can continue their daily lives or operate in the second economy for private gain,

perhaps engaging in small acts of 'everyday resistance' to express their displeasure in undetectable ways (Scott 1985).

Alternatively, they can attempt to leave the DPRK. In a tightly controlled and surveilled authoritarian context such as North Korea or the former German Democratic Republic (GDR or East Germany) individual 'exit' may be the only realistic option to overtly resist (Mueller 1999). Since the 1990s tens of thousands of North Koreans have defected from the DPRK. Today over 30,000 North Koreans reside in South Korea while perhaps as many or more live in China.

This paper proposes an agent-based model to simulate the dynamics and patterns of 'exit' from North Korea since the early 1990s. In a context like North Korea with few sources of reliable information computer modeling can help explicate and clarify how socio-political processes may work. The model relies on two simple core assumptions about people's decision to exit. First, people consider their relative prospects for economic success if they leave. If their current situation in North Korea is acceptable and if they do not expect things to improve by leaving, then they will be less likely to exit. Conversely if their current situation is dire and if they expect a major improvement in their financial prospects upon leaving then they will be more likely to exit. Second, people consider the likelihood that they will be caught and punished by the state. If their expectation of repression is low then all else being equal they will be more likely to attempt to exit while if their expectation of being repressed is high then they will be less likely to attempt defection. The model takes into account that these factors vary over time and across space in North Korea, using empirical input into the simulations.

The next section provides an overview of defection from North Korea. After then explaining the model specifications in more detail the paper engages in a simulation to better understand the drivers and dynamics of 'exit' from North Korea. Following the simulation, the model and its

results are compared to evidence from a variety of sources to explore its utility and validate its findings. The paper concludes with remarks about the utility and limitations of the model, its applicability to other contexts, and the theoretical insights it provides.

Exit from North Korea: Background

Prior to the 1990s defection from North Korea was extremely rare. While North Korea's relationship with the rest of the communist world before that time ensured that there were some interactions and travel within the communist bloc (see Armstrong 2013), defection to the capitalist world was rare and was often for specific political reasons. Prior to 1994 the number of North Korean defectors entering South Korea could be counted in the dozens. This changed with the famine that devastated North Korea during the mid-1990s (see *Table 1*).

Table 1: Number of North Koreans Arriving in South Korea by Year

Year	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
	8	8	52	41	56	86	71	148	312	583	1139	1281	1894
Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**TOTAL**
	1383	2091	2554	2803	2914	2402	2706	1502	1514	1397	1275	1418	30212*

*71% Female; 29% Male

The famine had its causes in the years during and after the collapse of the Soviet Union as preferential economic arrangements with the rest of the communist world drastically vanished during the early 1990s (see Haggard & Noland 2007). During and after the famine at least three processes helped contribute to increased levels of defection from North Korea. First, state capacity was crippled during the famine. This meant that the Public Distribution System (PDS), the DPRK's rationing scheme for allocating food and other necessities, could not provide for the vast majority of the population. Other state institutions like schools, factories, and even security services were working at severely diminished capacity and thus could not enforce dictates from Pyongyang consistently.

Second, given the lack of food and the breakdown in state capacity, people resorted to "self help" to find the nutrition necessary to survive (Smith 2009). This meant either turning to technically illegal 'second economy' activities like bartering and selling goods and services, scavenging public lands or forests for food in lieu of attending school or work, or indeed attempting to escape from North Korea entirely (Joo 2010). Third, North Korea's neighboring countries – China and South Korea – had been growing economically at some of the world's highest rates for the preceding decades. This meant that the 'push' factor of North Korea's crumbling economy may have been exacerbated by the 'pull' factor of the possibility of a better life in China or South Korea (see Green 2016).

As a result of these processes defection from North Korea skyrocketed, but from a low starting poitn. As *Table 1* demonstrates, the numbers of North Koreans entering South Korea before the year 2000 was under five hundred total. After the year 2001 that number never dropped below 1,100 annually meaning that there are now over 30,000 North Koreans in South Korea. For comparison, however, it is worth noting that the proportional numbers of those defecting from

North Korea remain low relative to another state divided by the Cold War: the GDR. Between the time that the Berlin Wall was erected in 1961 and demolished in 1989 the lowest number of East Germans leaving as refugees was 3,512. When combined with other categories of outmigrants the lowest total in a given year was 11,343 and the highest was 42,632 out of an East German population of about 16 to 19 million (see Hirschman 1993). North Korea's lower numbers of defections and higher population (about 23 to 24 million) mean that exit from the DPRK is still rare relative to perhaps the most similarly situated Cold War creation.

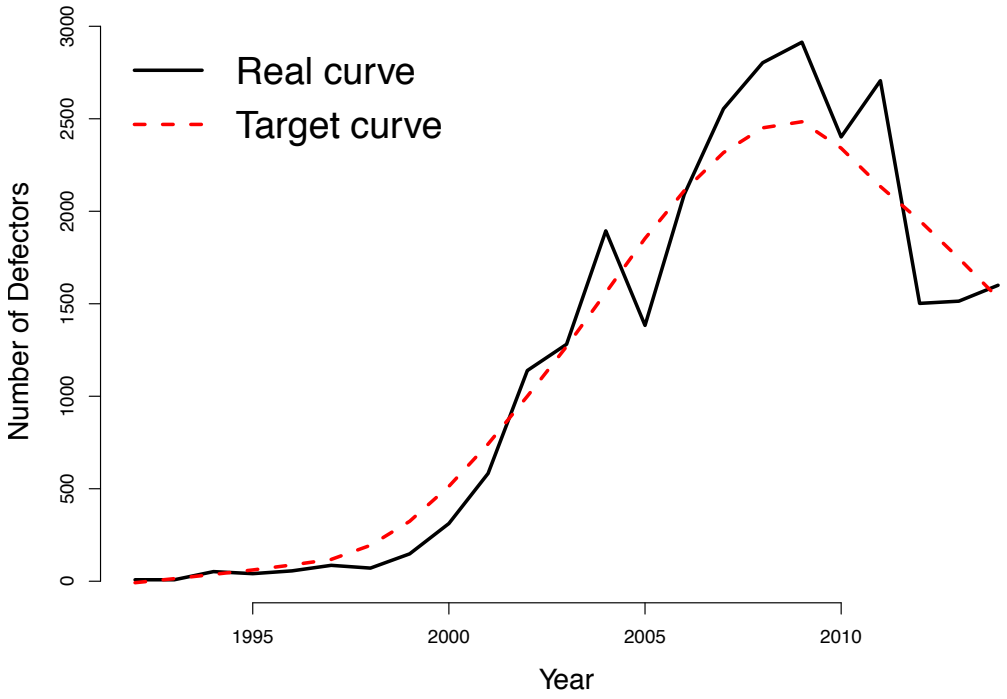

Figure 1 Estimated defection levels by year ("real curve") and target trend for simulations ("target curve").

Since the numbers in *Table 1* only capture those who made it to South Korea, however, two points should be noted. First, there is a time lag between when North Koreans leave the DPRK and when they make it to South Korea. Most North Koreans escape into China first, but since their presence is illegal in China they must make it to a third country such as Mongolia or Thailand in order to present themselves to a South Korean consulate to be eligible to go to the South. In a survey of 300 North Korean defectors in South Korea, Haggard and Noland (2011) found that 44% of respondents spent more than three years in China between the time they left North Korea and arrived in South Korea. This means that of those in *Table 1* who arrived in South Korea in 2002, for example, almost half would have left North Korea in the mid-to-late 1990s.

Second, many North Koreans who defect never make it to South Korea either because they cannot or choose not to. The numbers of North Koreans residing in China are very unreliable not only because their presence is illegal in China but also because there are several hundred thousand ethnic Koreans who are citizens of China living in areas near the border with North Korea and among whom North Koreans may be able to blend. Estimates of North Koreans living in China range from 30,000 to 300,000 while one research effort documents a decline from a high of roughly 75,000 in 1998 to about 10,000 in 2009 (Robinson 2010). The authors of this research attribute the decline to "tighter border security, increased migration to South Korea and other countries, and lower expectations of what is available in China" (ibid.)

Modelling defection behaviour

The starting point of the model design is a standard diffusion process, where individuals within North Korea emulate behavior of other defectors, such that the defection of some leads to the

defection of others. Similar to Kuran's (1995) or Lohmann's (1994) models of participation in pro-democratic protests in Eastern Europe, we would expect the participation of a few to affect the risk calculations of subsequent potential defectors. Both models are based on Granovetter's (1978) cascading revolutions, where each individual has a baseline propensity to participate and will only act upon this propensity when levels of participation correspond to their baseline threshold. In other words, someone who is strongly inclined to participate will do so, but someone who is moderately inclined will only participate when observing a small number of existing participants.

This model could in theory be applied to the North Korean context. Some individuals will have a very strong inclination to defect and do so regardless, while others will have a slightly less strong inclination and only defect when they see at least a few successful examples. Observing the success of some will slightly reduce their estimation of risk involved and therefore stimulate them to try. It is in principle possible that the distribution of preferences is such that a very small proportion of the population is inclined to defect, while by far most have a threshold value that is a lot higher, such that the cascade stops early. It is also possible that an event such as the famine reduced the threshold for some, thus triggering the cascade.

There are reasons why this model does not suffice in explaining the levels of defection in North Korea, however. While there is a cascade visible with increasing numbers of defections taking the shape of the first part of an S-curve, famous in diffusion studies (Rogers 1995), the increase in defections is extremely slow. While about 50 individuals would defect around 1995, this increased to around 2500 in 2007, which as a proportion of the population is an increase from just one in 500,000 to one in 10,000. While these increases are significant, they are hardly a cascade. More importantly, after 2011 the trend clearly reverses and numbers are going down

significantly again, from about one in 8,000 in 2009 to 1 on 15,000 in 2014. All these numbers are based on rough estimations, but the decline is clearly visible. The cascading model of revolution only explains developments in one direction, not its reversal, but an understanding of the dynamics of defections in North Korea means modelling both.

In our model we focus similarly on a baseline propensity to defect – which we model by demographic characteristics of the individual as well as changes in the economic circumstances in the country – and an expected level of repression, which is updated based on observed defections. Successful defections therefore lead to an updates estimation of the level of repression and thus a lowering of the threshold to defect, with a similar linear correlation between baseline propensity and the threshold. However, where the model diverges from the cascading revolution model is that the update of the expected level of repression can also be in reverse, whereby observing a failed attempt at defection increases the threshold to defect. We can therefore model not only the start, but also the decline of a cascade. Furthermore, the potential defector is not the only agent in the model, but the political regime also adjusts investment levels in repression of defectors based on observed levels of defection. So not only expected levels of repression change, also actual levels are modified over time.

The behavior of the state in terms of repression levels is based on the idea of a safety valve for discontent (Hirschmann 1993; Hoffmann 2005). The argument begins from the premise that the primary objective of the state is to remain in power and to do so it will repress threats (Davenport 2007). However because repression can entail political costs and risk backlash it is often seen as a strategy of last resort (Josua and Edel 2015). In some circumstances this may incentivize a minimal amount of defection. Those with the highest levels of discontent might cause more disruption to the regime when staying in North Korea – by persuading others of their anti-regime

views, for example – than when they leave. Allowing potential troublemakers to defect can therefore act as a safety valve, reducing pressure on the state to repress. There is some evidence that this occurred in North Korea around 1999 or 2000 as Kim Jong Il "apparently issued instructions that those who showed that they only went to China for food and work should be treated with a degree of leniency" (United Nations 2014: para. 386). The regime's challenge is to minimize resources spent on repression from defection while maintaining a minimal flow of defection as a safety valve – which will have to be constantly adjusted in light of changing circumstances.

The main challenge in our modeling approach is the extremely small proportions of the population that actually defect. While the numbers of not insignificant and increasing over time, we are still considering only one in 8,000 people who in any given year successfully defect. Furthermore, if we only model annual events, our simulation would contain too few iterations to draw any reasonable conclusions about the fit of the model with the empirical data. We therefore opt for a finer level of granularity by modelling weekly levels of defections between 1992 and 2014, which means that in any given iteration of the simulation at most 50 out of a population of nearly 25 million individual citizens defect. Since any model that incorporates only a very small number of individual agents, as is the norm due to computational limitations, would therefore be a very unrealistic model of the actual proportions, we instead develop a model of the entire population of the country, directly facing the challenge of finding interesting dynamics, while remaining at extremely low levels of defection.

Simulation implementation

Model setup

The agents in this agent-based simulation consist of two sets: the approximately 25 million individual citizens of North Korea[2] and the ten provinces (Ryanggang, North Hamgyong, South Hamgyong, Kangwon, Jagang, North Phyongan, South Phyongan, North Hwanghae, South Hwanghae and Pyongyang), which each have individual levels of repression. It is rare for agent-based simulations to model an entire population such as this – although there is a growing literature on synthetic populations – and it is computationally intensive to individually model the behavior of each agent and store the current state of each. However, in our simulation model there is very little variation between individual agents, given their location of residence and a number of demographic variables. We therefore store information on each agent only for each unique combination of demographic variables, keeping track of just the number of individuals in each category.

For each (category) of agent we estimate the utility of defection, which we arbitrarily scale from zero to one, and which is based on a combination of the motivation to defect and the ability to defect. Different circumstances will provide better opportunities for some demographic groups, e.g. people who live closer to the border, or have jobs from which departure will not be noticed as quickly, while others will have lower abilities to defect even if they want to. Similarly, circumstances will affect the need to defect, whereby the famine is of course the most important trigger, but also general economic well-being or welfare within the North Korean regime.

[2] In our simulations, we have 20.8 million agents at the start of the simulation and 24.5 million at the end, following empirical population data in North Korea from 1992 to 2014.

The demographic variables we include are selected on the basis of existing information about the demographics of defectors, relative to the overall population (e.g. Haggard and Noland 2011; DPRK Census 2008). Where we know there are a disproportional number of defectors among a particular category, this is a relevant variable to include in our model. Gender is the first variable that is included, based on information that over 70% of the defectors who make it to South Korea are women (see *Table 1*). The finding among qualitative researchers is that women are often better able to leave their place of work, as they are less often in government employment (Joo 2010). We also include age, where the assumption is that older people are less, and younger people more, able to defect if they wish to do so because of the physical demands associated with the rugged terrain in the border areas (Fahy 2015: 133). Somewhat correlated with gender, people who work in the secondary economy, rather than the primary, state-driven economy, have better opportunities to defect, but are perhaps less motivated to do so, as their economic circumstances will be better (Dukalskis 2016). An important feature of North Korean society is that people are divided in different status levels, from high status to low status, which impacts on both the ability and the motivation to leave (Collins 2012). Citizens in provinces bordering China are assumed to have much better opportunities (ability) to defect than citizens who live deeper in North Korea or closer to the nearly impenetrable demilitarized zone with between the DPRK and South Korea (Haggard and Noland 2011).

Aside from the demographic characteristics, we have two more variables that impact on both the ability and the motivation to leave, which are variables that are common across all agents but vary over time. These are a dummy variable to indicate the famine years, where in years of famine, citizens will have significantly higher levels of motivation to defect, and a measure of the overall economy, where we take annual growth rates in North Korea for the years we

simulate. *Figure 2* provides an overview of all demographic and economic variables we use as input to our model, where the relative levels of impact on ability and motivation are inserted by assumption, based on rough estimates from qualitative research.

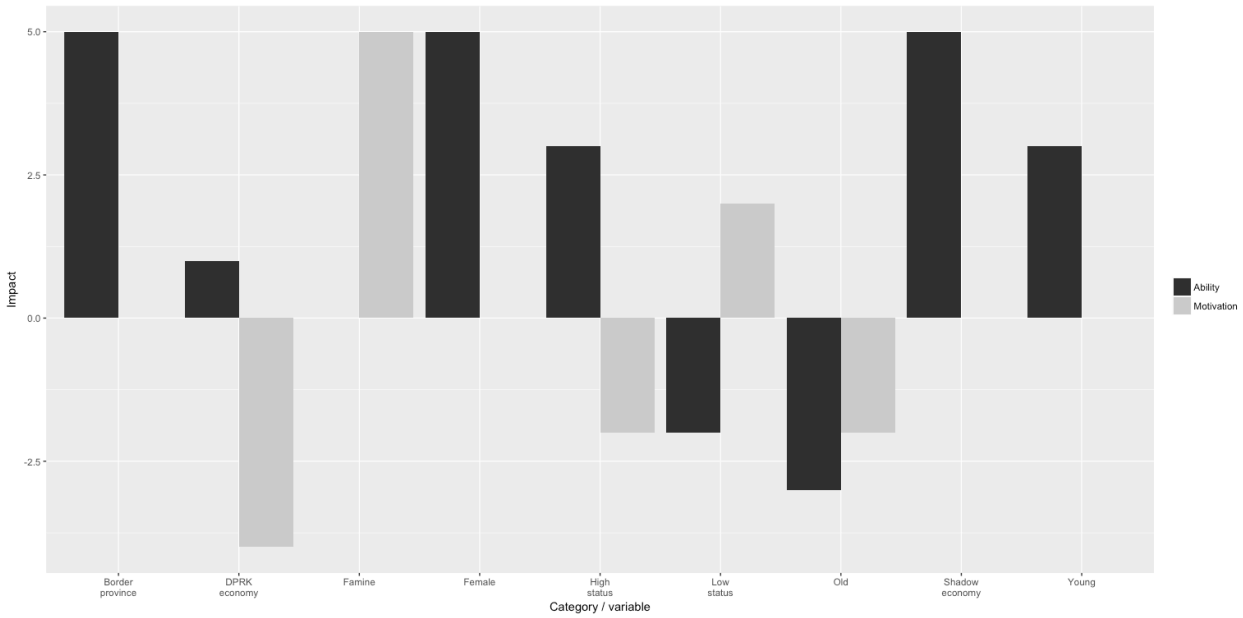

Figure 2 Assumed impacts of variables on motivation and ability to defect.

The starting point of the simulation is therefore that we have 720 different combinations of demographic categories, each representing a proportion of the populations, which is where available based on North Korean census data, but otherwise assumed as a more or less equal distribution. For each agent we calculate a baseline utility of defection and each agent as an expectation with regards the level of repression that would take place if he or she decided to defect. The state agent has an initial level of repression. These variables are then updated in each iteration over the 1,196 weeks that we simulate in our study.

Model behaviour

The starting point of the simulation is therefore that we have 720 different combinations of demographic categories, each representing a proportion of the populations. These proportions are based where available on North Korean census data from 2008 and otherwise assumed as a more or less equal distribution. The proportions are fixed throughout the simulation. For each agent we calculate a baseline utility of defection and each agent has an expectation with regards the level of repression that would take place if he or she decided to defect. The state agent has an initial level of repression. These variables are then updated in each iteration over the 1,196 weeks that we simulate in our study. Each iteration (i.e. week), a certain proportion of each category will decide to defect and these defections will be counteracted with a certain level of repression. Based on the relative success of these defections, agents will then update their expectations with regards the repression levels, and provinces will adjust their levels of repression. Figure 3 provides a schematic overview of the different simulation steps each iteration.

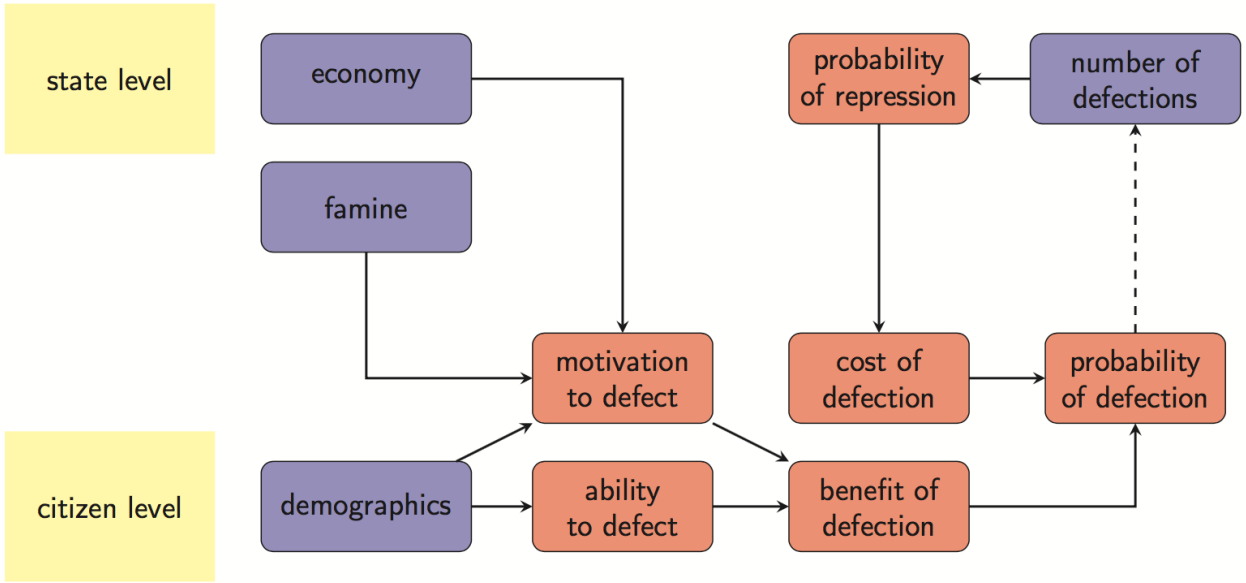

Figure 3 Overview of simulation flow in each iteration. The dotted line represents an aggregation, solid lines represent factors in each step. Blue boxes are province-level characteristics, red boxes are variables at actor level, both citizens and province-level government (repression rates).

Updating expectations of repression by the citizen

Repression levels are measured as probabilities. There is a certain probability, unknown to the citizen, of the state repressing – or rather, succeeding to repress – a defection. While theoretically the state has both a motivation to have a probability below one because of the safety valve argument, and an ability lower than one because they will not be able to avoid every single defection, we do not make a distinction between these two aspects in our modeling strategy. Therefore, the actual level of repression can be denoted by a single number, namely the probability for any given defection that the state will be able to stop this defection. We assume that all repression is binary – either the defection succeeds or the defection is avoided – while we do not model the fact that repression can be partial or subsequent to defection. Often, repression takes the form of punishing the family of a defector that stays behind (Hawk 2012). Here, however, we model this phenomenon only in the sense that a defector might decide not to defect because of this, or might succeed in defecting, leaving other nuances outside of the model. Since the actual level of repression is a probability, it makes sense to think of the expected level of repression as an estimate of this probability, with a certain level of uncertainty. As more defections and potential repressions are observed, the citizen will update this view, reducing the level of uncertainty, and converging on the actual repression rates. An appropriate probability distribution to model such estimates and associated uncertainty is the beta-distribution, which is a probability distribution over the support from zero to one, thus appropriate for modeling probabilities that are also bounded to this range. The beta-distribution has two shape parameters, and α and β, say, whereby higher values of both reduce the variance, higher values of α correspond with higher level of expected repression and vice versa for β. Updating the prior

value of a beta distribution given a set of new, known attempts at defection, is straightforward: $\alpha_1 = \alpha_0 + N_R$ and $\beta_1 = \beta_0 + N_D - N_R$, with N_D the number of known defections and N_R the number of those that were repressed. In our simulations, we include parameters that restrict the level of visibility of defections, as not all attempts at defection will be visible. We parameterize the number of weeks a citizen can look back to observe defections and repressions, and we parameterize the probability of each single attempt at defection to be observed.

Updating levels of repression by the state

The updating of the level of repression by the state is straightforward. We parameterize for each simulation the maximum number of repressions that are tolerated within a province and the speed by which a state adjusts its level of repression based on observed repressions. Unlike citizens, we assume that the state observes all defections. Thus, if in any given week the level of defections is higher than the tolerated amount, the level of repression is increased by a fixed percentage, and if this is not the case, it is reduced by the same percentage.[3] The level of repression is encoded as the probability that defection will be repressed, $Pr(repress) = R_{pt}$, at at any time $t = 1, 2, \cdots, T$, where $p = 1, 2, \cdots, N_p$ denotes the province. Note that we take a very binary view of repression: either there is repression and the defection fails, or there is no repression and the defection succeeds.[4]

[3] The results in this working paper are based on this simple model of state behavior, but we also experiment with more complicated models, whereby the level of motivation to defect among citizens is taken into account – the safety valve should be larger if levels of discontent are high – or where we take into account qualitative evidence regarding levels of state repression in North Korea over time.

[4] An alternative, presumably, would be to think in terms of varying levels of repression, resulting in varying levels of cost to the defector.

Determining the probability of defection of each citizen

We are trying to model the probability of defection, $Pr(defect) = D_{it}$, for each individual agent $n = 1, 2, \cdots, N$. This will be based on two elements: the utility a particular individual derives from successful defection, U_{it}^{D}, and the expected level of repression, $ER_{\{p[i],t\}}$, based on the observed history of repression. This estimated level of repression does not vary across individuals within a province, i.e., $E_{it} = E_{p[i],t}$. We assume that each defection, whether repressed or not, has a fixed probability of being known to other agents, $Pr(known) = K$, and that past defection that took place more than a fixed time period H ago is forgotten. K and H are therefore fixed, exogenous parameters to the model.

The utility of defection for any given agent – or rather, category of agent based on demographic characteristics – is updated each iteration based the demographic characteristics, economic conditions, and the approximate levels of impact as visualized in Figure 2. The benefit or utility of defection is based on using logistic regression formulations to predict first the level of ability, then the level of motivation, and finally the utility as an interaction model between the two:

$$ability_{it} = (1 + \exp\big(-s(3young_{it} - 3old_{it} + 5female_{it} + 5secondary_{it} - 2lowStatus_{it} + 3highStatus_{it} + 5borderProvince_{i} + 1economy_{t})\big))^{-1},$$

whereby s is a simulation parameter adjusting the overall impact of variables on ability, while the relative impacts of each variable are fixed. Similarly for motivation:

$$motivation_{it} = (1 + \exp\big(-\delta famine_{t} - s(4economy_{t} - 2old_{it} - 2rural_{it} - 2lowStatus_{it} + 2highStatus_{it})\big))^{-1}.$$

We then model the utility or benefit of defecting as:

$$U_{it}(d|\neg r) = (1 + \exp(-(\gamma_1 + \gamma_2 motivation_{it} \cdot ability_{it})))^{-1}.$$

Given a particular expected utility for an individual citizen, the expected level of repression is taken into account to assess the overall pay-off of repression. Symbolically, $U_{it}(d|\neg r)$ is the utility for individual i at time t derived from defection given that this defection is not repressed. The symbol $d|\neg r$ denotes defection which is not being successfully repressed, whereas $d|r$ denotes defection that is successfully repressed by the state. We assume $U_{it}(d|r) = 0$. The expected pay-off of defection is then given by

$$E_{it}\left[P_{\{p[i],t\}}\right] = \left(1 - E_{it}\left[R_{\{p[i],t\}}\right]\right) \cdot U_{it}(d|\neg r) + E_{it}\left[R_{\{p[i],t\}}\right] \cdot C,$$

where C denotes the loss of being repressed (and consequently being sanctioned).[5] We assume $C = -1$ for any agent. This guarantees that the loss of being repressed is higher than utility that one can obtain from defection. Thus, the expected payoff becomes

$$E_{it}\left[P_{\{p[i],t\}}\right] = \left(1 - E_{it}\left[R_{\{p[i],t\}}\right]\right) \cdot U_{it}(d|\neg r) - E_{it}\left[R_{\{p[i],t\}}\right].$$

Accordingly, the probability of defection is given by

$$Pr\left(\left(1 - E_{it}\left[R_{\{p[i],t\}}\right]\right) \cdot U_{it}(d|\neg r) - E_{it}\left[R_{\{p[i],t\}}\right] > 0\right) = Pr\left(E_{it}\left[R_{\{p[i],t\}}\right] < \frac{U_{it}(d|\neg r)}{1+U_{it}(d|\neg r)}\right).$$

As $E_{it}[R_{p[i],t}]$ follows a beta-distribution, we take the cumulative distribution evaluated at $\frac{U_{it}(d|\neg r)}{1+U_{it}(d|\neg r)}$ as the probability of defection.

[5] Note that the expected level of repression is here represented by a probability distribution reflecting the level of uncertainty about the level. This is therefore not an expectation in the statistical sense of an expected value.

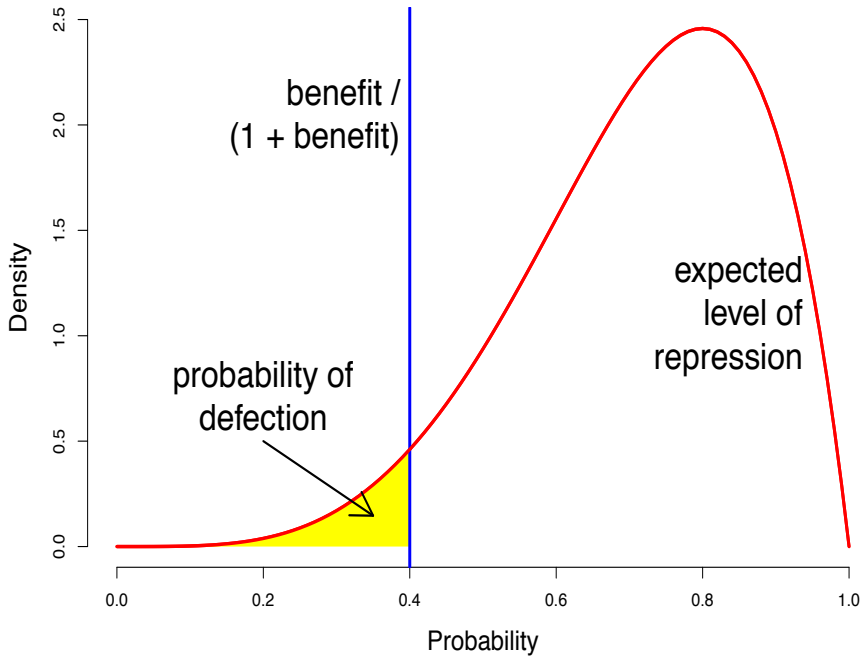

Figure 4 Updating the probability of defection based on expected utility and

Figure 4 then outlines how this level of benefit is related to the expectations regarding the level of repression to determine the probability of defection. The red line denotes the beta-distribution of expected levels of repression, in this example with relatively high levels of uncertainty, such that the variance of the beta-distribution is high. As attempts at defection take place and are observed – depending on the above mentioned visibility parameters – the level of uncertainty reduces. The probability of defection is then a function of the expected benefits and the expected level of repression as depicted in Figure 4, the probability being the cumulative probability, or area under the beta probability density function, up to benefit. Thus, as expected levels of

repression increase, this surface reduces, and as the level of benefit increases, this surface is larger.

Altogether the parameters of the simulation model are therefore as follows:

- Initial (prior) shape parameters of the beta distribution (α_0 and β_0).
- Coefficients of the benefit function (γ_1 and γ_2).
- Maximum tolerated level of defection and speed of adjustment of repression rates.
- Initial repression rate.
- Probability that citizens observe a defection or repression (K) and the time frame which they observe (H).
- The impact of the famine (δ) and the overall impact of variables on the ability and motivation to defect (s).

Together this generates a rather large parameter space within which we search for the combination of parameters that approaches the curve in Figure 1 closest, which given the extremely low empirical probabilities of defection is a genuine challenge.

Results

There is a significant challenge in searching for a combination of parameter settings whereby the defection rate remains extremely low, yet shows the kind of curve we observe – i.e. one where there are nearly zero defectors around 1992, which then leads to a small cascade as the famine hits, but which declines again as it hits about 3,000 defectors per year – in a model with a

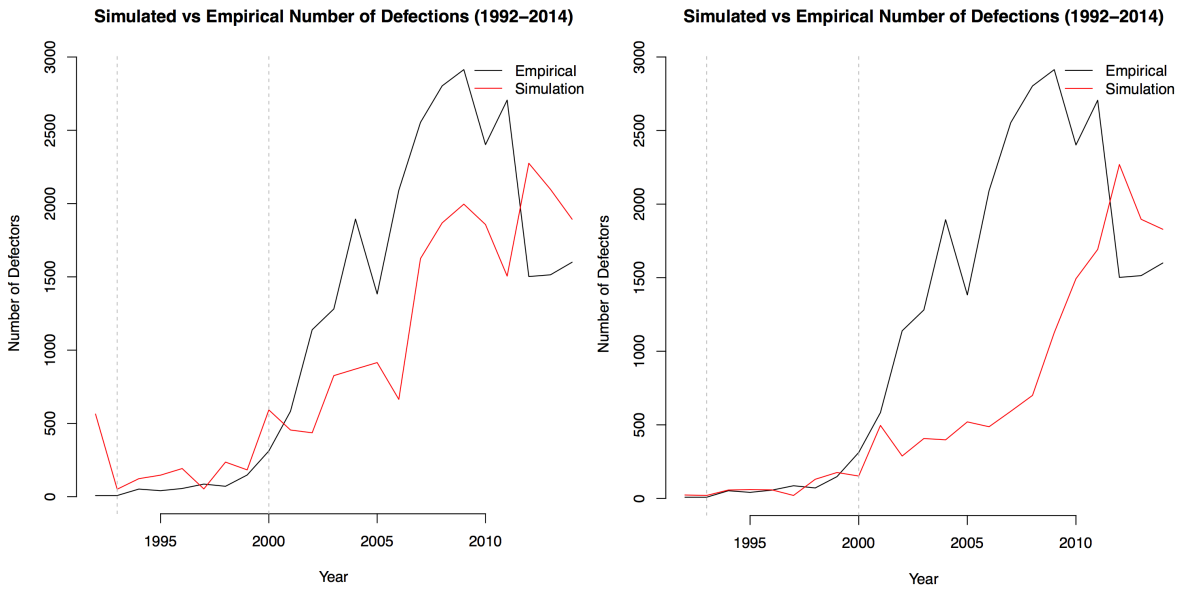

Figure 5 Two of the best results obtained, showing similar trends and levels to the empirical data.

relatively high number of parameters. Under most configurations, there are either no defections – because agents are insufficiently motivated or repression rates are too high – or there is an immediate unrealistically high stream of defectors, or a cascade is triggered which immediately expands to too many defectors. In the high dimensional space defined by the initial parameters of the model, the subset where we obtain reasonable results is very small indeed, and therefore difficult to identify.

In this work in progress version of the paper, we have obtained good results in a very small number of simulations, and have therefore some idea where this is located. These results are preliminary, however, as this also means that we have not fully worked out the exact boundaries of that subspace, nor can we be certain that this is the only region in the parameter space where we obtain such results. These simulations were found using an algorithm inspired by the Metropolis-Hasting algorithm in Markov Chain Monte Carlo analysis (Gamerman and Lopes 2006: 128-129), where we calculate a score for each simulation reflecting the match with the

empirical data,[6] and after each simulation propose new parameter values that are randomly selected close to the previous simulation, and whether or not to continue the parameter search in that direction depends on the relative improvement in the score – if it improves we continue, if it does not, we randomly decide whether to continue or not. Figure 5 provides a graphical depiction of two of the best results obtained thus far. In terms of the overall score, the figure on the left is the highest, due to the close tracking of the empirical trend data. The figure on the right, however, shows greater similarity in terms of the very first start, where the famine triggers the first defections – an important qualitative difference that makes it align more closely with the North Korean narrative.

Because it is unlikely that there is a clear linear relationship between parameters and simulation output, we use the machine learning technique of random forests (Hastie, Tibshirani and Friedman 2009: ch 15) to identify the parameters of key importance to obtaining good results – results are presented in the appendix. The two primary factors turn out to be the rate of adjustment of repression and the impact of the famine (δ). While we vary δ between approximately 0 and 2,[7] all high scoring simulations have a δ of 0.84 on average with a standard deviation of 0.04. The impact of the famine is thus modest – otherwise the start of the famine immediately triggers a large cascade of defections – but enough to initiate a small cascade. The adjustment of the repression rate is based on a percentage increase or decrease when the level of defection is above or below the tolerated level by the provincial government. Simulations with

[6] This score is largely based on the vertical distances between the empirical and the simulation series, but also awards points for running a longer simulation and for having at least some defectors. Many simulations have a very low defection rate which suddenly spikes. These simulations tend to obtain good scores in terms of *average* deviation from the line, even though the pattern is not similar to the empirical data. Since simulations automatically stop as defections cross a preset threshold, we take into account whether this stopping mechanism was triggered or not in calculating the score.

[7] This range is only this short after "zooming in" on the region with better results.

fast adjustment do not generate results similar to those in the empirical data, as they do not allow the citizens to sufficiently learn from previous defections and adapt their behavior. We obtain higher scores only when repression rates are adjusted by 0.1% or less, more typically 0.05%. Note, however, that the simulation ticks are weeks, so these repression rates are adjusted on a weekly basis. On an annual basis, an adjustment rate of 0.05% amounts to an annual adjustment of 30% if defection rates remain higher than tolerated throughout the year.

The importance measure of the variable from the random forest analysis drops significantly for the remainder of the parameters. The initial values of the beta distribution that represents prior perceptions of repression is the next most important. Here it is the relationship between α_0 and β_0 that matters, where we obtain good scores when $\alpha_0 = 26 + 0.66\,\beta_0$, with a residual standard error of only 4.01. In other words, there is a strong linear relationship and α_0 needs to be higher than β_0, such that the expected value of the beta distribution remains in the same region – while across simulations this varies from 0 to 1, the full spectrum of the beta distribution, for the simulations where we obtain high scores, the expected value of the expected level of repression is between 0.47 and 0.65, and scores remain very low outside of this range. In other words, potential defectors early in 1992 expect a fifty-fifty chance that their defection will succeed.

A variable of similar importance to the initial values of the expected level of expression is H, the number of weeks that potential defectors look back to assess the level of repression. This is a key variable of the diffusion mechanism built into the model, as this reflection on previous level of defection and repression affects the decision whether or not to defect. In our simulations, this variable ranges from about 200 to 1,196, the total number of weeks in the simulation. In other words, if $H = 1196$ then there is a full memory, any potential defector considers all previous defections in the simulation before deciding on a course of action. We find some reasonable

scores across the range, but the higher scores all occur around a memory of about 380 weeks, with a standard deviation of 53. So we again see a relatively narrow region in parameter space where good results are obtained in the simulations, and we do not obtain any good simulations when potential defectors either learn too little or too much from the past. The intuition behind this result is that when one learns from too many past defections, the precision of the beta distribution is too high, and the potential defector can precisely assess the level of repression. This means that the area under the curve in Figure 4 becomes very small indeed. On the other hand, when memory is too short, there is no diffusion effect, and we will not obtain the cascade that is visible in Figure 1. Related, and somewhat less important in predicting which simulation will give good results, is the probability that a defection will be observed. Here we note that all high scoring simulations have a probability of observation (K) of around 0.50, with a standard deviation of 0.044.

The last important variable worth mentioning is the initial level of repression at the start of the simulation. While simulations vary more, with good scores obtained across a range of initial values for this parameter, the best results are obtained when the initial repression rate is at least 0.8, i.e. there is an 80% chance, or higher, that any defection will be repressed. While according to the random forest analysis this variable is not key in identifying successful simulations, the tolerated level of defection, the safety valve, also shows a clear pattern: while evaluated on a range from about 100 to about 500, all higher scoring simulations have a mean tolerance level of 180 defectors in the province in a given week, with a standard deviation of only 5.64.

Interpretation, validation and limitations

Through these results we gather a picture of the configuration where the simulations lead to levels of defection that are somewhat aligned with our empirical observations. At the start of

the simulation, citizens are relatively unclear about the level of repression, giving themselves a 50% chance of success if they decide to defect. This is optimistic, as the actual success rate is less than 20%. As initial defections take place, this perception will be gradually updated and become more realistic, but there are limitations to this process: each defection (or repression) has only a 50% chance of being observed, and anything that happened more than seven years ago is forgotten. When defections deviate from the tolerated level of defection – the 180 defections that are tolerated from within the province to ensure a safety valve of discontent – the government adjusts repression levels, but this adjustment is relatively slow.[8] Finally, the famine impact variable is clearly important for the results, which suggests that this is indeed what triggers the cascade that we observe in the simulations and that matches the empirical data.

The results thus far are remarkably clear, in that the region currently identified in the space of initial parameters where we obtain good results is very small indeed, with good simulation results occurring within relatively tight bounds on a range of different and substantively interesting parameters. Since we know very little of developments inside North Korea, where we cannot directly observe failed defection attempt or changes in repression levels and provincial priorities and where we do not have direct survey data on levels of discontent, ability to defect, or perceptions of repression levels, and only have rough indications of aggregate levels of defection, it is of interest that a simulation that makes relatively few assumptions about how defection decision are being made only resembles the empirical data in very specific circumstances. This is not as robust as a statistical analysis of empirical observations on our

[8] In the two simulations reported in Figure 5, defection levels never reach the tolerated level and repression rates continually decline, from about 85% to about 44% at the end of the simulation. By that time the beta distribution of expected levels of repression is a relatively tight distribution around this value.

explanatory variables, since it relies more heavily on a range of modelling assumptions, but it is highly suggestive.

These simulations have some limitations, however, that further iterations of the paper will need to address. First of all, while we currently validate on the basis of aggregate statistics, there are some more detailed demographics available on defectors that can be used to validate not only the total number of defections, but also who defects. Does the model predict the right groups in the population to defect? The next important improvement would be to separate the coefficient on the economy variable. Currently this measures the impact of changes in the growth rate of the gross domestic product on the motivation to defect and the level of impact is based on the assumed relative impacts depicted in Figure 2, but Figure 2 is primarily based on knowledge about demographics of defectors, which has little to do with aggregate changes in the economy. Since one of our primary variables of interest is the impact of the famine on defection and repression levels, we will need to separate the impact of economic developments more generally, and therefore allow this to be a free parameter in the model specification. Finally, more generally, the sensitivity to the specification in Figure 2 needs to be investigated.

Conclusion

North Korea is perhaps the most impenetrable country in the world, where there is very little empirical data available for political science research. For our understanding of the functioning of authoritarian regimes this case study is of particular interest, however, exactly due to its unique levels of constraints on information both internally and externally, and its high levels of repression and control over cross-border interaction. One of the key characteristics of the regime is that it puts severe limits on citizens leaving the country. In the simulation study presented here we investigate the dynamics of this repression, and the defection behaviour of

discontented citizens. Since empirical data on key variables of interest – individual level perceptions of repression, of the economy, and of opportunities inside and outside the country, and regime level variables on repression levels and the allocation of resources to protecting the borders – are unavailable, simulation is an alternative method to look into the black box. If we can build a simulation based on a few realistic assumptions about the behaviour of individuals and the state, we can investigate under what conditions we find similar levels of defection, and a similar trend since the beginning of the famine, from North Korea.

For the individual behaviour we model the level of motivation and ability to defect on the basis of demographics and economic circumstances, as well as a diffusion mechanism whereby individuals assess the levels of risk based on previous attempts to defect. For the regime we assume a safety valve policy, whereby the regime represses sufficiently to keep defection numbers low, but enough to let the most discontented citizens leave instead of contributing to resistance inside the country – to promote the exit option rather than the voice option, in Hirschman's (1993) terms.

The simulation results suggest that, given these modelling assumptions and the empirical trend in successful defections, the North Korean regime has indeed been reducing its repression levels over time, but also that originally individual North Koreans were relatively optimistic about their chances of success and that based on observed defection attempts, they have updated their expectations to more accurately reflect the actual levels of repression. It suggests that citizens are aware of about half the attempts at defection – the remainder goes unnoticed – and that they base their assessment on defection attempts over the previous seven or so years.

Future iterations of the paper will provide additional robustness checks on model assumptions, separate more clearly overall economic developments from the famine, and more carefully

evaluate whether the parameter space where we currently observe optimal results is indeed the only region – current results based on close to 100,000 simulations certainly suggest this is the case. Finally, the validation part will pay more close attention to the demographics of the successful defectors.

Bibliography

Armstrong, Charles K. (2013a) *Tyranny of the Weak: North Korea and the World, 1950-1992.* Ithaca: Cornell University Press.

Collins, Robert (2012) *Marked For Life: Songbun, North Korea's Social Classification System*. Washington DC: Committee for Human Rights in North Korea.

Davenport, Christian (2007) State Repression and Political Order. *Annual Review of Political Science* 10: 1-23.

DPRK Census (2008) *2008 Population Census: National Report*. Pyongyang: Central Bureau of Statistics, DPR Korea.

Dukalskis, Alexander (2017) *The Authoritarian Public Sphere: Legitimation and Autocratic Power in North Korea, Burma, and China*. New York: Routledge.

Dukalskis, Alexander (2016) North Korea's Shadow Economy: A Force for Authoritarian Resilience or Corrosion? *Europe-Asia Studies* 68(3): 487-507.

Fahy, Sandra (2015) *Marching through Suffering: Loss and Survival in North Korea*. New York: Columbia University Press.

Gamerman, Dani, and Hedibert F. Lopes. 2006. *Markov Chain Monte Carlo. Stochastic simulation for Bayesian inference*. 2nd edition. Boca Raton, FL: Chapman & Hall.

Granovetter, Mark. 1978. "Threshold models of collective behavior." *American Journal of Sociology* 83: 1420–1443.

Green, Christopher (2016) The Sino-North Korean Border Economy: Money and Power Relations in North Korea. *Asian Perspective* 40(3): 415-434.

Haggard, Stephan & Marcus Noland (2007) *Famine in North Korea: Markets, Aid and Reform*. New York: Columbia University Press.

Hassig, Ralph and Kongdan Oh (2009) *The Hidden People of North Korea: Everyday Life in the Hermit Kingdom*. Lanham, MD: Rowman & Littlefield.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009), *The Elements of Statistical Learning. Data mining, inference, and prediction*. 2nd edition, Springer.

Hawk, David R. (2012) *The Hidden Gulag: The Lives and Voices of 'Those who are Sent to the Mountains'.* Washington, DC: US Committee for Human Rights in North Korea.

Hirschman, Albert O. (1993) Exit, Voice, and the Fate of the German Democratic Republic: An Essay in Conceptual History. *World Politics* 45(2): 173-202.

Hoffmann, Bert (2005) Emigration and Regime Stability: The Persistence of Cuban Socialism. *Journal of Communist Studies and Transitional Politics* 21(4): 436-461.

Joo, Hyung-min (2010) Visualizing the invisible hands: the shadow economy in North Korea. *Economy and Society* 39(1): 110-145.

Josua, Maria and Mirjam Edel (2015) To Repress or Not to Repress: Regime Survival Strategies in the Arab Spring. *Terrorism and Political Violence* 27(2): 289-309.
Kuran, Timur. 1995. Private truths, public lies. The social consequences of preference falsification. Cambridge, Massachusetts: Harvard University Press.
Lohmann, Susanne. 1994. "The dynamics of informational cascades: the Monday demonstrations in Leipzig, East Germany, 1989-91." *World Politics* 47(1): 42–101.
Mueller, Carol (1999) Escape from the GDR, 1961-1989: Hybrid Exit Repertoires in a Disintegrating Leninist Regime. *American Journal of Sociology* 105(3): 697-735.
Robinson, Courtland (2010) *North Korea: Migration Patterns and Prospects*. NAPSNet Special Reports 4 November 2010. Available at: http://nautilus.org/napsnet/napsnet-special-reports/north-korea-migration-patterns-and-prospects/
Rogers, Everett M. 1995. *Diffusion of innovations*. 4th ed. New York: The Free Press.
Scott, James C. (1985) *Weapons of the Weak: Everyday Forms of Peasant Resistance*. New Haven: Yale University Press.
Smith, Hazel (2009) North Korea: Market Opportunity, Poverty and the Provinces. *New Political Economy* 14(2): 231-256.
United Nations (2014) Report of the Commission of Inquiry on Human Rights in the Democratic People's Republic of Korea.

Appendix: random forest results

	importance
Rate of adjustment of repression rates	9747806.7
Impact of famine	8607270.7
Length of memory of past defections	2892008.6
Alpha parameter of prior beta distribution	2604975.6
Initial repression rate	2414993.1
Probability any defection is observed	1912530.7
Typical impact of demographic variable	1788773.7
Logistic coefficient on interaction between motivation and ability	1279852.8
Beta parameter of prior beta distribution	1087050.5
Maximum tolerated level of defection	929615.1