Bayesian Population Projections for Every City in the World

Leontine Alkema, Jennifer L Chunn,
Patrick Gerland, Danan Gu and Gerhard K. Heilig[1]

September 23, 2011

[1]Author's affiliations: L. Alkema and J. L. Chunn, National University of Singapore, Singapore. P. Gerland, D. Gu and G. K. Heilig, United Nations Population Division, Population Estimates and Projection Section, New York City, New York, USA. Correspondence to Leontine Alkema, email: alkema@nus.edu.sg. The project described was supported by a grant from the Global Asia Institute at the National University of Singapore. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Global Asia Institute. Also, the views expressed in this paper are those of the authors and do not necessarily reflect the views of the United Nations. Its contents have not been formally edited and cleared by the United Nations. The designations employed and the presentation of material in this paper do not imply the expression of any opinion whatsoever on the part of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Abstract

The United Nations Population Division publishes the World Urbanization Prospects, which includes projections for all capital and major cities in the world. A deterministic model is used to construct the city projections, based on the most recent observed growth differential between the city population and the urban population in the country. An uncertainty assessment of future outcomes is lacking. In this paper, we propose a Bayesian time series model to capture changes in city-urban growth differentials over time, and to construct probabilistic city projections. We evaluated the Bayesian city projection model by an out-of-sample exercise, and found that the proposed model is reasonably well calibrated and has smaller projection errors than the current projection method.

1 Introduction

While the growth of the world's population has been slowing down in recent decades, the urban population has been growing very rapidly (from 220 million to 2.8 billion) over the 20th century, and has passed a milestone in 2009 with more than half of the human population, 3.5 billion people, living in urban areas. At the global level, all future population growth will be in towns and cities. Already today, more than 38 per cent of the world urban population live in 442 cities of more than 1 million people, and the UN projects that about 100 more cities will reach 1 million within the coming decade (United Nations, 2010).

While urbanization - the increase in the urban share of total population - is somewhat inevitable and goes hand in hand with development, the spatial distribution of the population, and especially its concentration in major urban areas offer not only great economic potentials but also important challenges for urban planning, housing and infrastructure, water supply and sanitation, provision of services, power and transportation, institutions and governance, and many other social, cultural and environmental aspects. Therefore, knowing better which urban areas are experiencing the fastest growth and whether this growth will continue or slow down in the future is of great concern to planners and decision makers responsible to design and implement policies aimed to insure a more equitable and sustainable development for future generations.

The urbanization process is estimated and projected in the World Urbanization Prospects (WUP) which is published by the United Nation Population Division and revised every two years. The most recent revision from 2009 (United Nations, 2010) includes estimates and projections for all capital cities and all cities with more than 750,000 residents, from 1950 until 2025. A deterministic model is used to construct the city projections, based on the most recent observed growth differential between the city population, and the urban population in the country. The UN method, however, does not provide any uncertainty assessment in the projections.

When asked to evaluate evaluate country population projections developed by various agencies (United Nations Population Division, the World Bank, and U.S. Census Bureau), Bongaarts & Bulatao (2000) concluded that the development and refinement of projection uncertainty should have "high priority" (p13). We believe this same sort of urgency is necessary when making population projections for cities. Developing projection intervals that have probabilities attached to them would help users of the projections evaluate alternative future situations, and the likelihood that they would occur. We also believe that with the development of any projection method its predictive ability should be evaluated. One such way is to withhold recent data, making projections on old data and comparing the projections with the withheld data. Furthermore, probabilistic projections should be shown to be accurately dispersed, e.g., a 80% interval should capture the truth 80% of the time.

In this paper, we propose a Bayesian city projection model (BCP) that can be used to project city populations for cities from all countries and of all sizes. Our projections do include uncertainty bounds, which we evaluate by an out-of-sample exercise. We do so by fitting the model to data available from 1950 to 1995, producing projections from 1995 to 2009 and assessing the accuracy of our point estimates (posterior median projection) and probabilistic projection intervals.

2 Methodology

2.1 Data

The UN maintains an internal database with data on over 5,000 cities from 230 countries or areas. City population time-series are collected from figures published by national sources, such as censuses and administrative registers. The database contains cities that have reached a population of over 100,000 at any instance between 1950 and 2009. It also monitors over 1,200 cities which have less than 100,000 residents.

Table 1 gives a summary of the number of observations per each city in the data base. For a vast majority of cities, there are no more than 5 or 6 observations of its city population.

Table 1: Number of observations in each city between 1950 to 2009.

Num Obs	Num Cities	Pct	Cumul Pct
1	147	2.9	2.9
2	405	8.0	11.0
3	835	16.6	27.6
4	575	11.4	39.0
5	566	11.2	50.2
6	1174	23.3	73.6
7-10	1076	21.4	95.0
$>$10	254	5.0	100.0

For computational ease, the initial results presented in this paper are based on a random subset of 200 cities of any size, which are located in 101 different countries. The total number of observations in this test data set is 1,341.

2.2 Model

The set-up of the Bayesian city projection model is motivated by the current UN city projection model. We first discuss the UN model, then explain our proposed projection model.

UN city projection method The UN city projections are based on the projections of the difference between the city growth rate and the growth rate of the urban population of the country that the city is located in. This difference in growth rates is called the city-urban growth differential, denoted by $r_{i,t}$:

$$r_{i,t} = c_{i,t} - u_{i,t}, \tag{1}$$

where $c_{i,t}$ is the city growth rate for city i in year t, and $u_{c,t}$ is its urban growth rate. The city-urban growth differential is illustrated in Figure 1, using the New York-Newark urban agglomeration, USA, as an example. New York-Newark's estimated population is given in Figure 1(a) (in blue), and its corresponding city growth rate is shown in Figure 1(b). The urban growth rate for the US is added to plot (b), and the difference between the city population growth rate and the urban population growth rate, thus the city-urban growth differential, is shown in figure (c).

In the UN projection model, the growth differential is projected to converge to the so-called *global norm line*, illustrated in figure (d). The global norm line gives the expected growth differential $r_{i,t}$ as a function of the city population size:

$$r_{i,t} = 0.017089 - 0.00144 \log(C_{i,t}), \tag{2}$$

where $C_{i,t}$ is the city population size at year t. The global norm line is based on an analysis of the relation between growth differentials and city population sizes; the growth differential was found to decrease with city population size. The regression equation was fitted to the data relative to 1,982 cities located in the 113 countries that had at least 2 million inhabitants in 1995. Based on the global norm line, the city growth rate is expected to be slightly lower than the urban growth rate when the city population is greater than about 145,000.

The city-urban growth differential is projected to converge linearly to the global norm in 20 years, as illustrated in Figure 1(d) for the New York-Newark urban agglomeration. Future city growth rates then follow from the difference between the projected growth differential and the projected urban growth rate within the country, and are used to project future city population. The projections for New York-Newark are added in Figure 1(a).

Bayesian city projection model The set-up of the Bayesian city projection model (BCP) is motivated by the UN projection model; like the UN, we propose to project the city-urban growth differential for each city. However, instead of projecting the convergence of the differential to the global norm line, we project convergence of the growth differential to zero, based on the assumption that cities will eventually grow at the same rate as the urban population as a whole. This is modeled with a first-order auto-regressive (AR(1)) time-series model:

$$r_{i,t} = \rho_{i,t} \cdot r_{i,t-1} + \varepsilon_{i,t}, \tag{3}$$

with autoregressive parameter $0 < \rho_{i,t} < 1$ and error terms $\varepsilon_{i,t}$, with $E(\varepsilon_{i,t}) = 0$. In this time series model, the growth differential depends on its previous value. It is expected to decrease towards zero if the current growth differential is positive, and to increase if the differential is negative. The expected increase or decrease depends on $\rho_{i,t}$. This is illustrated in Figure 2, which shows three simulated trajectories of an AR(1) model. The starting value of the growth differential, $r_{i,1970}$, and the error terms $\varepsilon_{i,t=1970,\ldots,2009}$ are the same in each of the trajectories. Autoregressive parameter $\rho_{i,t}$ is held constant over time, and differs between the three trajectories. The figure shows that the trajectory with the largest ρ_i ($\rho_i = 0.75$) deviates from zero for longer than the other trajectories. For projecting city growth rates, this implies that the rate of convergence of a city growth rate to the urban growth rate will

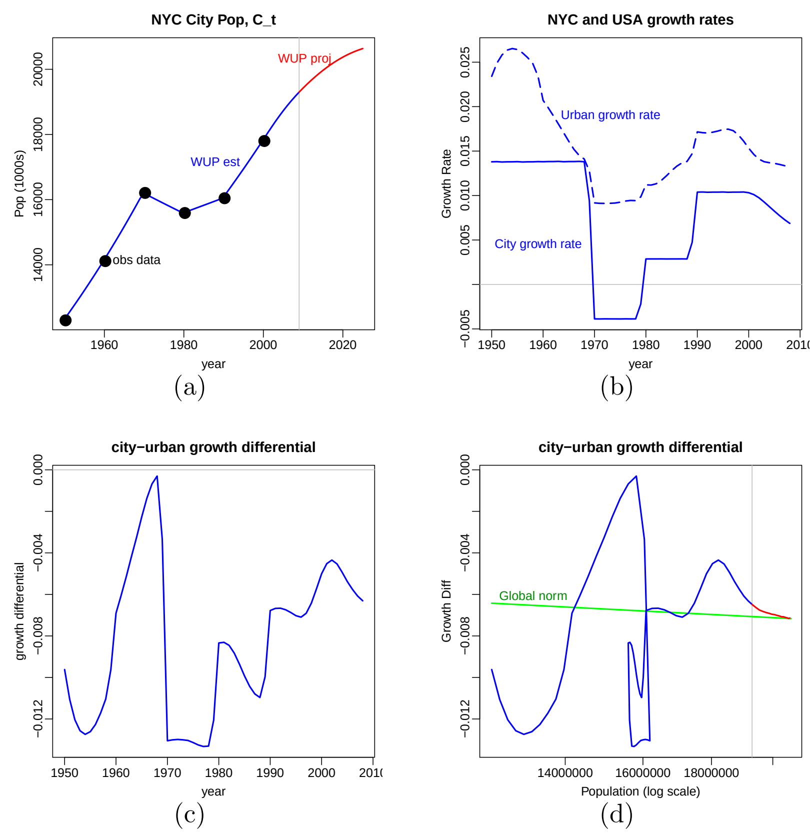


Figure 1: Illustration of the UN city projection method for New York-Newark urban agglomeration(NYC), USA. (a) Estimates and projections of the city population in NYC. (b) Observed city growth rate in NYC and urban growth rates in the USA. (c) Observed city-urban growth differential, (d) Convergence of city-growth differential to the global norm line.

depend on the current growth differential, as well as the values of $\rho_{i,t}$ in the projected period: the larger the $\rho_{i,t}$'s, the longer the period that a city is expected to be higher or lower than the urbanization rate within the country. The rate of convergence in the BCP is not fixed at 20 years as in the UN model, but will depend on the current growth differential and the autoregressive parameters $\rho_{i,t}$ during the projection period.

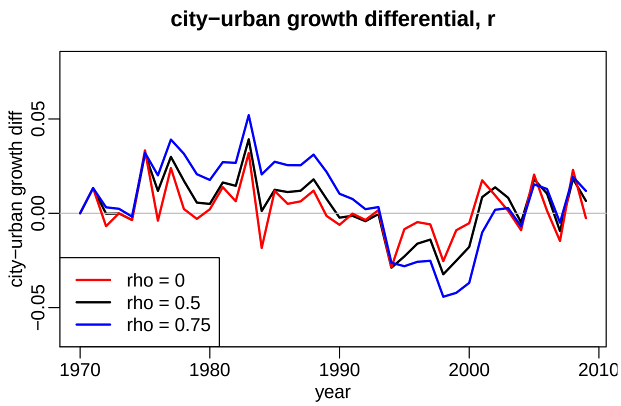


Figure 2: Illustration of the effect of different values of ρ on the fluctuations in the city-urban growth differential.

Our modeling approach for $\rho_{i,t}$ is based on the hypothesis that the rate of convergence of the city growth rate to the urban growth rate depends on the urbanization rate within the country. When a country's urbanization process is occurring most rapidly, city growth rates are expected to converge more slowly to the urban growth rate, since a city is likely to either be leading the national urbanization process or "catching up". On the other hand, if a country is at the start or at the end of its urban transition, cities and urban populations are more likely to grow at similar rates, and differences between a city's growth rate and the urban growth rate are less likely to sustain over time. We empirically evaluated this hypothesis by examining the relationship between the $\rho_{i,t}$'s and the urban growth rates, $u_{i,t}$ in the past in the cities in our data set. To do so, we first used the quintiles of the distribution of all observed urban growth rates during the observation period to classify the urban growth rates into five groups (1 = "very low", 2 = "low", 3 = "medium", 4 = "high" and 5 = "very high" urban growth rates). We then fit the time series model (3) to the observed growth differentials, estimating a different $\rho^{(j)}, j = 1, \ldots, 5$ in each "urbanization rate" group, e.g. $\rho_{i,t} = \rho^{(1)}$ if $u_{i,t}$ was in the "very low" urban growth rate group in year t, $\rho_{i,t} = \rho^{(2)}$ if $u_{i,t}$ was in the "low" urban growth rate group, etc. Figure 3 shows the estimates of $\rho^{(j)}$ for each group of urbanization rates. The result confirms our hypothesis that $\rho_{i,t}$ increases with $u_{i,t}$. We implemented this relation in the BCP, and let $\rho_{i,t}$ increase as a logistic function of urban growth rate, $u_{i,t}$,

$$\text{logit}\,(\rho_{i,t}) = a + b \times u_{i,t}, \text{or} \tag{4}$$

$$\rho_{i,t} = \frac{1}{1 + exp(-(a + b \times u_{i,t}))}, \tag{5}$$

where parameters a and b determine the shape of the logistic function. This function, as opposed to a piecewise function, allows for a smooth and parsimonious relationship between $\rho_{i,t}$ and $u_{i,t}$ with the necessity to only estimate two parameters. The estimated logistic fit of the function is added to Figure 3.

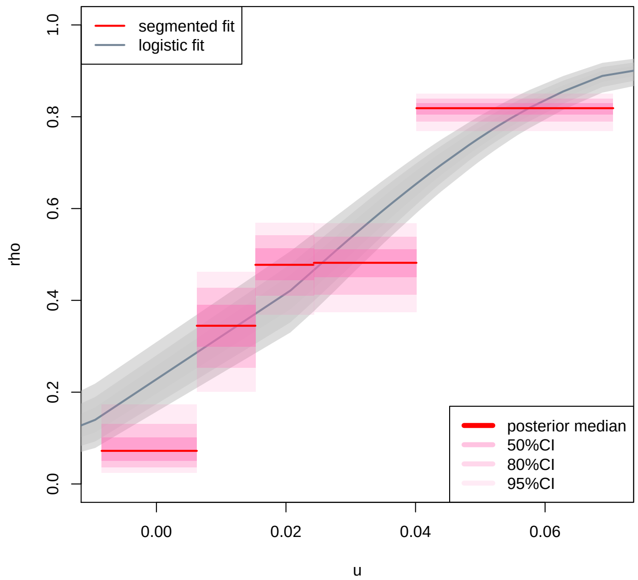


Figure 3: Posterior estimates of $\rho_{i,t}$ varying as a function of urban growth rate, $u_{i,t}$.

Given the time series model (3), the future trajectories of the city-urban growth differential are determined by the $\rho_{i,t}$'s, as well as the random distortions $\varepsilon_{i,t}$. The distribution of the random distortions is given by:

$$\varepsilon_{i,t} \sim t_{\chi}\left(0, \sigma_{\varepsilon}^{2}\right),$$

which is a t-distribution with χ degrees of freedom, mean 0 and variance σ_{ε}^{2}. The t-distribution is chosen, as compared to a normal distribution, to allow for heavier tails (a larger proportion of distortion terms that are far from the zero mean). The tail heaviness is dependent on the degrees of freedom χ, which is estimated from the observed growth differentials in the past.

Data model The time series model as specified above pertains to annual city-urban growth differentials. However, we do not have annual data. Instead, we observe average city-urban growth differentials, $y_{i,k}$ over an observation period of $T_{i,k}$ years, for $k = 1, \ldots, n_i$, where n_i is the number of observed growth differentials for city i. We assume that these observed growth differentials vary normally around the "true" average growth differential,

$$y_{i,k} \sim N\left(\bar{r}_{i,k}, \frac{\sigma_\delta^2}{T_k}\right),$$

where

$$\bar{r}_{i,k} = \frac{1}{T_k} \sum_{t=(\sum_{j=1}^{k-1} T_j)+1}^{\sum_{j=1}^{k} T_j} r_{i,t}, \text{ for } k = 1, \ldots, n_i,$$

and σ_δ^2 is the error variance of an observed growth differential over a 1-year observation period.

Prior distributions Diffuse prior distributions are assigned to the model parameters. Priors for both of the variance parameters are uniform priors with the bounds of the uniform chosen to be much wider than the posterior ($\sigma_\delta \sim U(0, 0.01)$ and $\sigma_\epsilon \sim U(0, 1)$). A uniform prior on the t-distribution degrees of freedom, ranging from 3 to 30 to ensure the variance of the distribution is not undefined. The prior for the city-urban growth differential in the year 1949 (represented by index $t = 0$) is given by the stationary distribution of the city-urban growth differential:

$$r_{i,0} \sim t_\chi\left(0, \frac{\sigma_\varepsilon^2}{1 - \rho_{i,0}^2}\right),$$

where $\rho_{i,0}$ is set to be the city-specific auto-regressive parameter in 1950.

We used an Markov Chain Monte Carlo (MCMC) algorithm to obtain samples of the posterior distributions of the model parameters, using the MCMC simulation program JAGS (Plummer, 2011). The "best" city projections are given by the posterior median of the projected city population, and the bounds of the projection intervals are given by the corresponding percentiles of that distribution.

3 Results

Because the model is still being refined, we refrain from discussing specific city projections in depth. Instead, we present a sample of city projections in Figures 4 and 5. For each city, we show the estimated and projected city-urban growth differential, city growth rate, and city population for both our BCP model and the UN model. The BCP results also include 50%, 80%, and 95% projection intervals.

The BCP median projections for the cities in Figures 4 and 5 are similar to the UN projections. In Figure 6, we present four cities that warrant further investigation. For Buffalo, USA, the BCP median projections are higher than the UN projections. The opposite holds true for Klang, Malaysia, where the BCP projects a smaller population size. For both cities however, the deterministic UN projections are within the BCP's 80% projection intervals. Monrovia, Liberia has had a non-stable growth history with the city population decreasing and increasing from observation to observation in the last 20 years, which lead to very different projections when comparing BCP to the UN. In Lilongwe, Malawi, the BCP median projections are essentially equivalent with the UN projections. But, the BCP projection intervals grow to be quite wide. While the interval sizes appear to be unnecessarily wide, this is not the case for all cities in the model. For all 200 cities combined, the width of the 80% projection interval is about $\pm 6\%$ after projecting for 5 years. This increases to 10.6% at 10 years and 14% at 15 years.

Cross-validation To evaluate the predictive ability of our current model, we assessed the coverage probabilities of the projection intervals and calibration of the median projections. To do so, we fit the model using all observations from 1950 to 1995 thereby creating a fictitious scenario where we are able to compare our projections with the true "future" city population observations from 1995 to 2009. We omitted 30.1% (413 out of 1341) of the observations for this cross-validation.

We also constructed deterministic projections using the UN methodology where future city-growth differentials converge to the global norm line, using the same observations. For certain cities, the UN performs post-hoc adjustments to city growth rates to ensure the cities do not have negative growth and to ensure the aggregated growth rate of all cities within the same country does not exceed the total urban growth rate. The adjustments only impact a small number of cities. For our cross-validation exercise, we do not implement these adjustments for any cities such that the BCP and UN projections are directly comparable.

Figures 7 and 8 show the out-of-sample projections for eight cities. The BCP median projections are shown in red with projections intervals in pink (50%, 80% and 95% PI) and the projections using UN methods are shown in blue. The observed city populations included in the model are depicted by green circles. The out-of-sample observations used as the "true future" populations are shown as green squares.

As shown in Table 2, our 80% projection intervals accurately included 74.5% of the "future" observations. To assess the accuracy of the projections, we assess our "best guess" projection, the median projection. Table 3 contains summary statistics of the projection errors, the difference between the "true" city population and the projected population by

BCP or UN method. For the BCP model, the median absolute error was 26.1, which means a typical BCP city projection was 26,000 people higher than the true observed city population. This was a 19% reduction from the median absolute error of the UN-method projections. Looking at just the projection error would be deceptive since an error in a projection of 10,000 is quite significant for a small city of 30,000 but trivial for a city larger than 10 million. As such, we look at the median absolute percent error, with the error scaled by the city size. The BCP median projection for a typical city was 6% of the "true" population compared to 8% median absolute percent error for the UN projections.

Expected coverage	Actual coverage
50.0%	56.2 %
80.0%	74.5 %
95.0%	87.4 %

Table 2: Cross-validation coverage probabilities.

	BCP		UN method	
	Mean	Median	Mean	Median
Error	-4.76	-0.93	6.65	-3.21
Abs Err	163	26.1	168	32.3
Sq Err	205,000	683	222,000	1044
Root Sq Err	453.3	26.1	471.0	32.3
Pct Error	-0.62	-1.04	-5.21	-2.22
Abs Pct Err	14.31	5.76	18.01	7.80

Table 3: City population cross-validation summary statistics for BCP and UN projections. The "Error" is the difference between the "true" city population and the projected population by BCP or UN method (in thousands). Also show are the absolute value of the errors ("Abs Err"), the squared value of the errors ("Sq Err"), and the square root of the mean/median squared error ("Root Sq Err"). Lastly, the errors (and absolute errors) are scaled by the "true" populations ("(Abs) Pct Err").

4 Discussion

We have proposed a Bayesian method to produce probabilistic projections of city populations for all cities in the world. Our out-of-sample exercise shows that the BCP model provides very reasonable projections. We are currently examining if the addition of demographic covariates into the model can help to produce more accurate projections and more narrow projection intervals. So far, we considered covariates such as the distance to the capital city, total population growth rate in the country, country total fertility rate and proportion of young adults in the country (age 15 to 24), but have not yet found a relationship between the city-urban growth differential and the covariates that has not already been accounted for.

Currently, the model has been fit for 200 random cities. We are working on extending the model to the 5,000 cities for which we have data. This lends itself to many potential logistical issues that are largely computational in nature. It will also allow for more exploration into developing a component of the model with more city-to-city correlations within the same country in addition to the urbanization rate.

Alternative city projections that include an uncertainty assessment were constructed by Montgomery and colleagues (Montgomery et al., 2009). In their city projection model, a projection of the city-specific total fertility rate is used as a predictor of future city populations. Rural-urban migration is not taken into account. A comparison of model performance between our proposed model and this existing modelling approach would be of interest but is not yet possible, because model validation of the alternative approach is lacking.

References

Bongaarts, J. & Bulatao, R. A. (2000). *Beyond Six Billion: Forecasting the World's Population*. Washington, D.C.: National Academy Press.

Montgomery, M. R., Buettner, T., Adamo, S. B., & Kim, D. (2009). Forecasting city population growth in develping countries. In *Proceedings of the 2009 Annual Meeting of the International Union for the Scientic Study of Population*, Marrakech, Morocco. http://iussp2009.princeton.edu/abstractViewer.aspx?submissionId=92468.

Plummer, M. (2011). JAGS v3.1.0: Just another gibbs sampler.

United Nations (2010). *World Urbanization Prospects: The 2009 Revision*. New York, NY: United Nations.

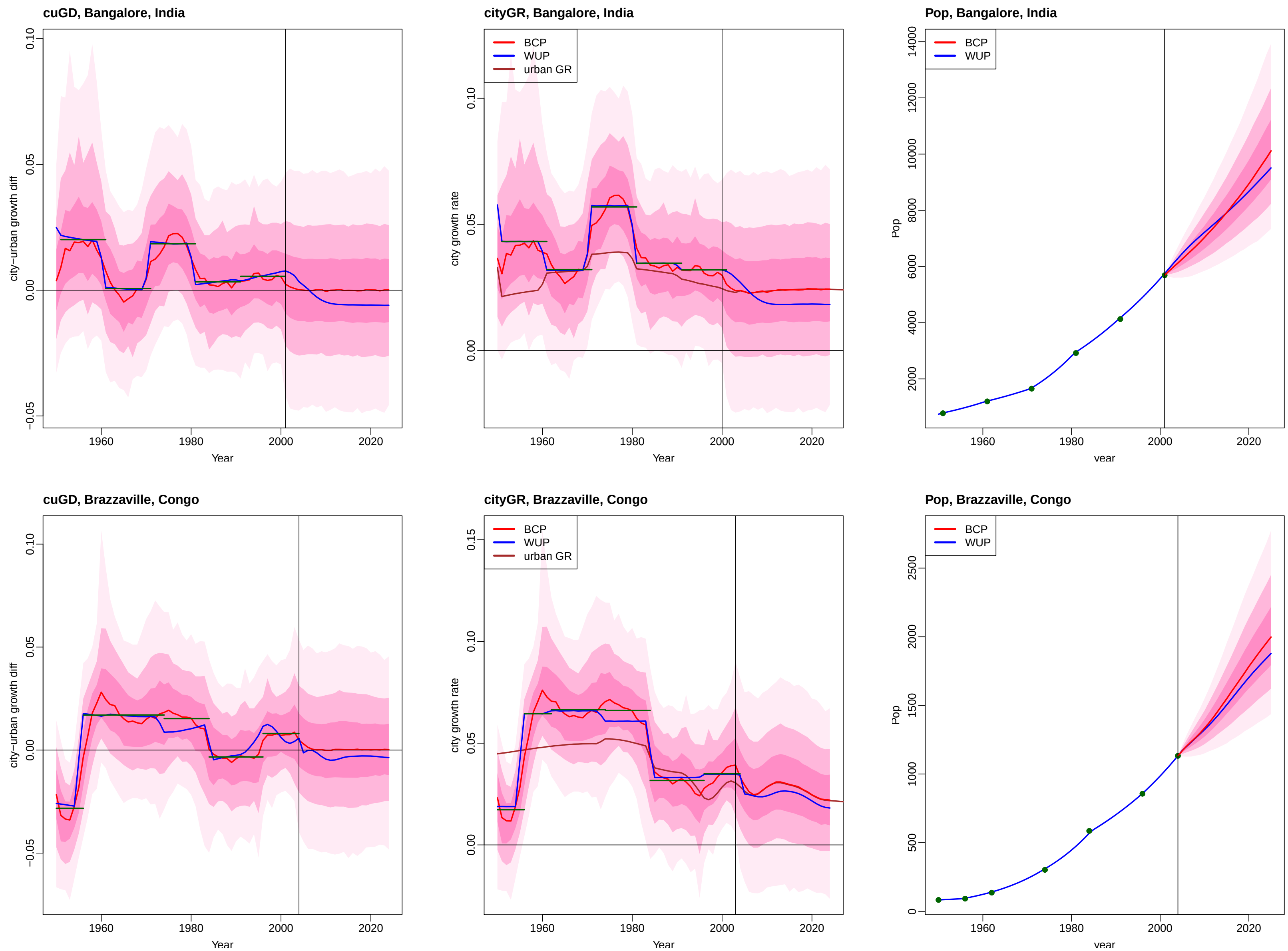


Figure 4: City population projections for Bangalore, India and Brazzaville, Congo.

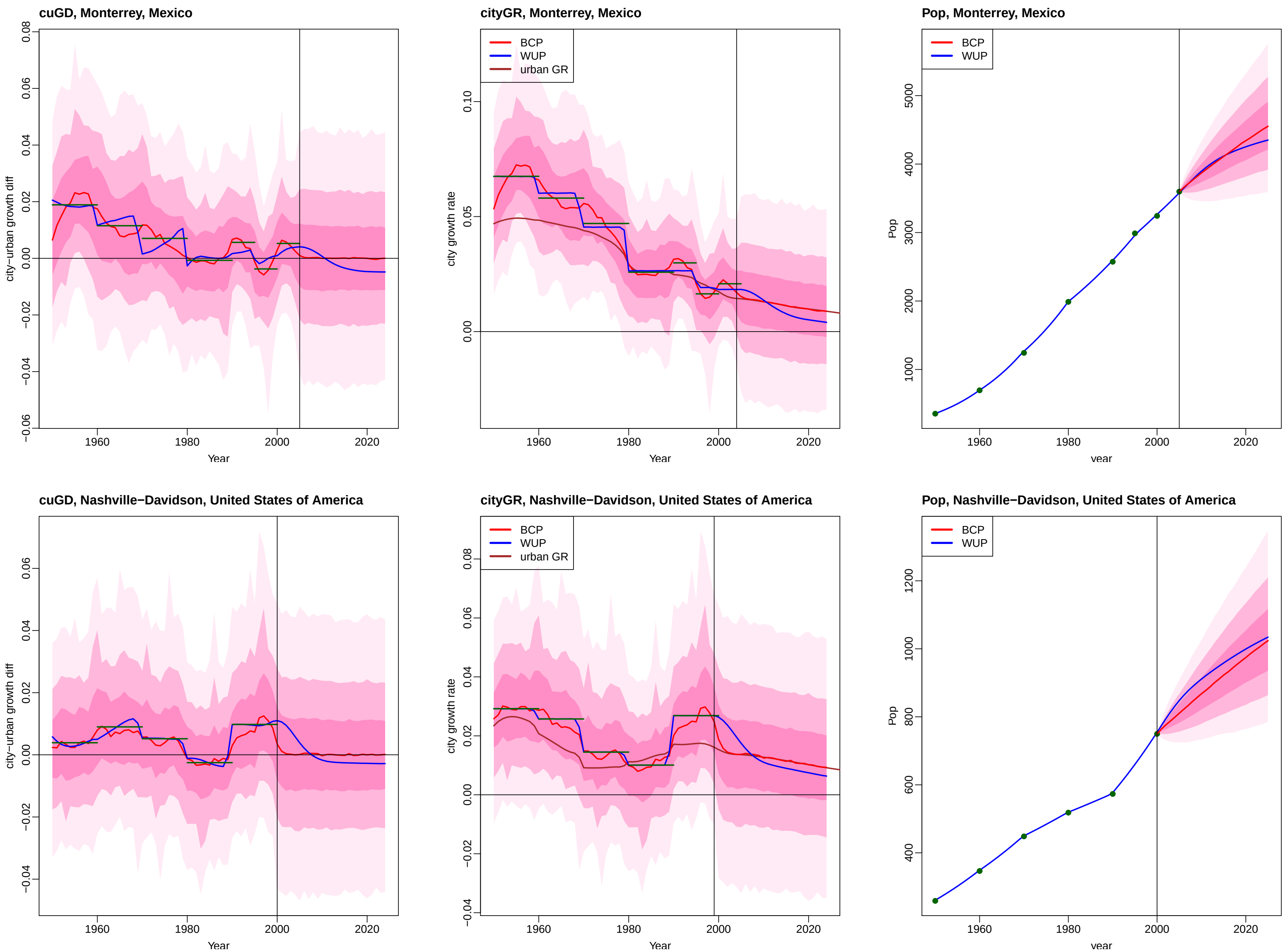
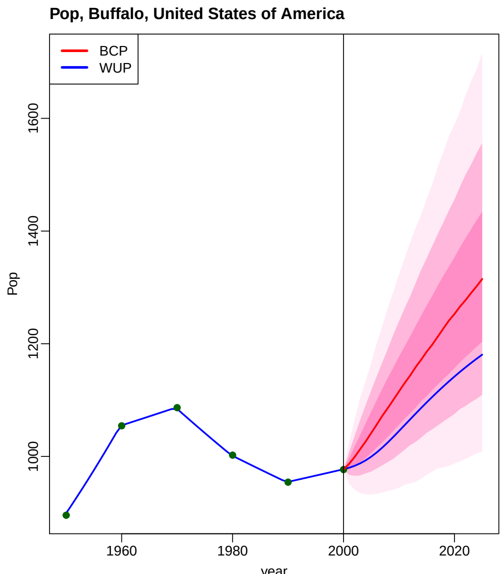
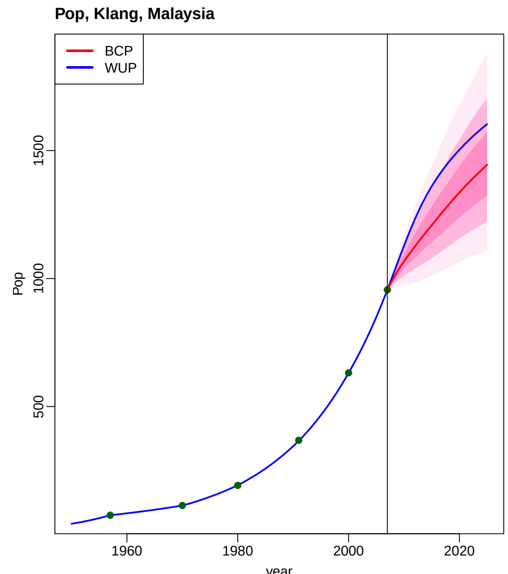


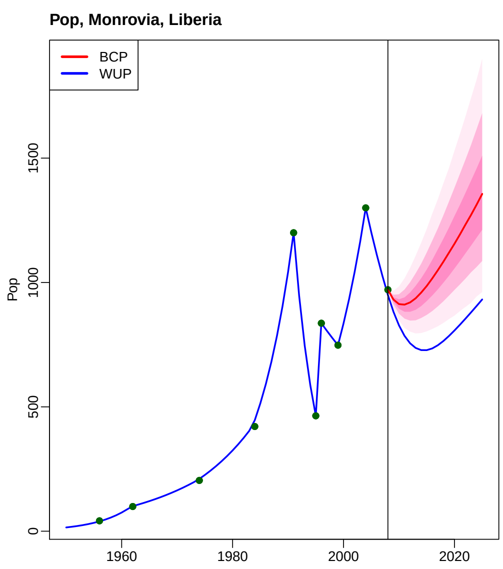
Figure 5: City population projections for Monterrey, Mexico and Nashville-Davidson urban agglomeration, USA.



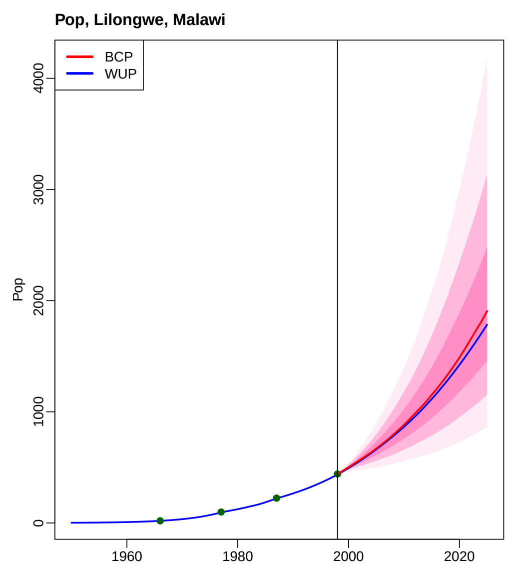
(a) BCP projections higher than WUP



(b) BCP projections lower than WUP



(c) Non-stable city growth history



(d) Large projection intervals

Figure 6: City population projections for Buffalo, USA; Klang, Malaysia; Monrovia, Liberia and Lilongwe, Malawi. These cities demonstrate some of the differences between BCP and the UN projections, an instance when the BCP projection intervals are very wide, and an "exception" case.

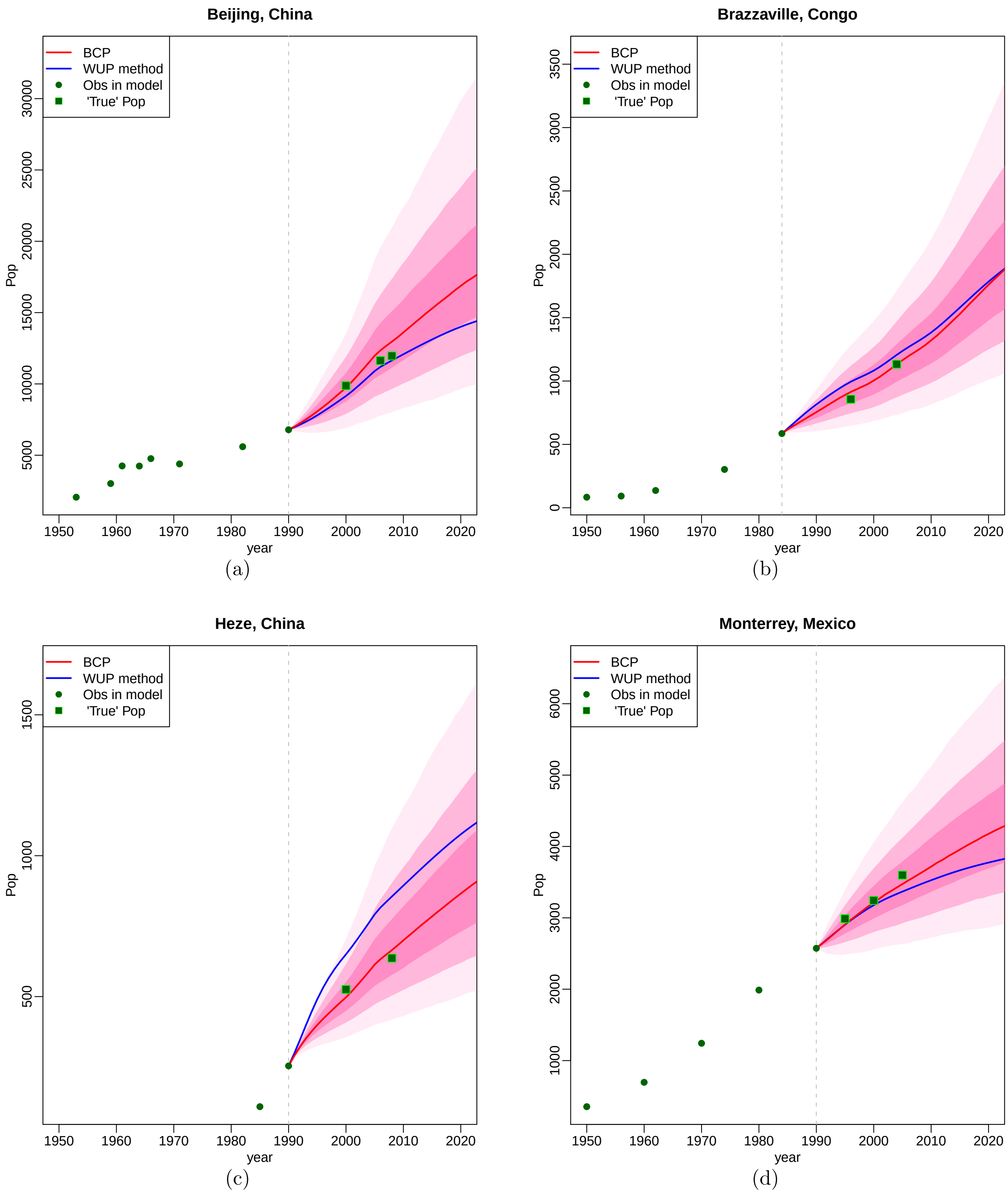


Figure 7: Examples of cross-validation.

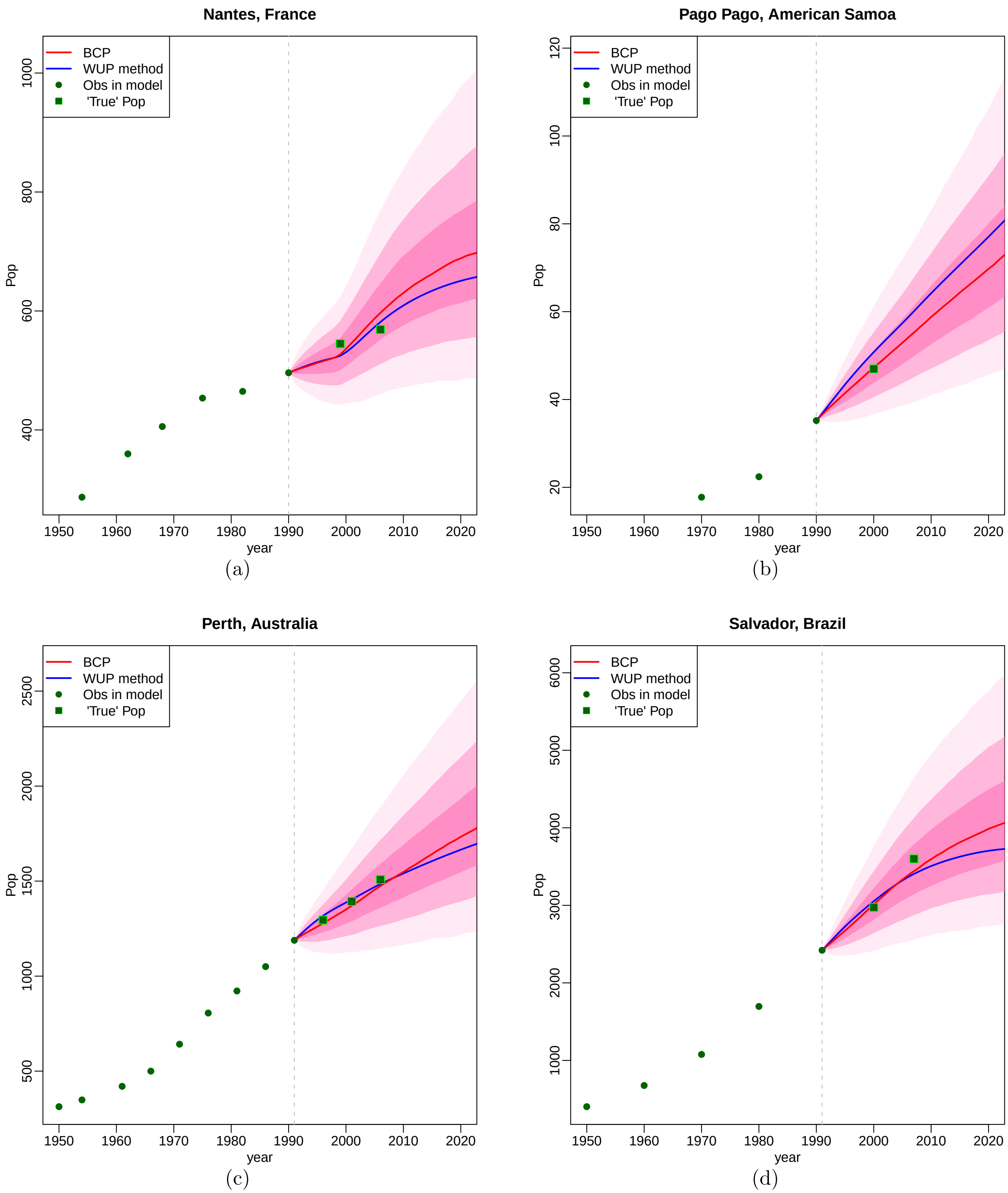


Figure 8: Examples of cross-validation.