

simPH: Tools for Simulating and Plotting Quantities of Interest from Cox Proportional Hazards Models

Christopher Ganrud

Yonsei University

Abstract

I describe an R package **simPH** for simulating and plotting quantities of interest estimated from Cox Proportional Hazards models. The quantities of interest include hazard ratios, relative hazards, first differences, marginal effects, and hazard rates for linear, linear interactive, time-varying, polynomial, and penalized spline estimated effects. The quantities of interest calculated from parameters drawn from the posterior, sampling distribution. The draws are then plotted using a methods that utilize the popular **ggplot2** package. The package is illustrated using epidemiological and political science data.

Keywords: cox proportional hazards, uncertainty, posterior distribution.

Results estimated from Cox Proportional Hazards (PH) models are often reported with coefficient point estimates or hazard ratios—exponentiated coefficients—and confidence intervals. However, this approach can (a) obscure the functional form of the estimate, especially over time and when it is non-linear or interactive and (b) give an inadequate sense of the uncertainty surrounding the estimates. There are currently limited capabilities in R¹ and popular statistical software such as Stata to present results in any other way. The **simPH** (Gandrud 2013) R package aims to make it much easier for researchers to show their Cox PH results.

Drawing directly from approaches taken by King, Tomz, and Wittenberg (2000) and Licht (2011) **simPH** simulates parameter from estimated Cox PH models by drawing them from their posterior distribution. Then it calculates quantities of interests, such as hazard ratios, first differences, relative hazards, marginal effects (for interactions), and hazard ratios. Finally, it plots the distributions of these simulated quantities of interest with visual-weighting (Hsiang 2012). The plots are generally created with the popular R package **ggplot2** (Wickham and Chang 2013) and as such can be aesthetically improved in virtually any way allowed by the package.

In this paper I discuss and provide examples of how to use **simPH** to simulate and plot quantities of interest estimated from Cox PH models. First, I discuss the quantities of interest **simPH** calculates and the simulation approach it uses to estimate uncertainty about these quantities. Then I discuss the general steps a researcher takes to use **simPH**. Finally, give step-by-step instructions for how to use **simPH** to show results from a variety of quantities of interest for a number of different kinds of estimated covariate relationships including linear, linear interactive, time-varying interactive, polynomial nonlinear, and penalized spline

¹The **survival** (Therneau 2013) and **Zelig** (Owen, Imai, King, and Lau 2013) packages do included limited functions for presenting estimates from Cox PH models with uncertainty. However, they have very limited or nonexistent capabilities to show nonlinear and interactive estimates with associated uncertainty.

nonlinear.

1. Simulating Quantities of Interest

Results from Cox PH models, and indeed many other methods, are often conveyed using ‘train-timetables’ of coefficient estimates and some measure of uncertainty such as standard errors or confidence intervals. This approach is inadequately informative, especially for effects that are estimated to vary over time, over values of the covariate, or in interaction with other covariates. When researchers do visually communicate uncertainty they often rely on graphical features, such as confidence interval lines that emphasize edges—areas of low likelihood—rather than the center, i.e. estimates that are more strongly supported by the model. The **simPH** package makes it easy to more fully present results from Cox PH models.

1.1. Calculating Quantities of Interest

First let’s look at the types of quantities of interest researchers are often interested in estimating from Cox Proportional Hazards models (Cox 1972). A basic Cox PH model is given by:

$$h(t|\mathbf{X}_i) = h_0(t)e^{(\beta\mathbf{X}_i)}. \quad (1)$$

$h(t|\mathbf{X}_i)$ is the hazard rate for a unit i at time t , i.e. the rate at which an event of interest—e.g. cancer remission, a war breaks out—happens. $h_0(t)$ is the baseline hazard at time t , i.e. the hazard when all covariates are 0. β is a vector of coefficients. \mathbf{X}_i is a vector of covariates for unit i .

Researchers are often interested in how a given covariate x affects the hazard rate $h(t)$. The simplest approach is to simply report the coefficient β for the covariate. It is more common to report the hazard ratio in its simplest form is e^β . It is a ratio in that it represents the ratio of the hazards of units with two levels of the covariate—all else equal—at a given point in time (t). In the case of e^β , the ratio is between a unit j with $x = 1$ compared to a unit l with $x = 0$. Generically for units j and l the hazard ratio is:

$$\frac{h_j(t)}{h_l(t)} = e^{\beta(x_j - x_l)}. \quad (2)$$

In the special case where $x_j \neq 0$ and $x_l = 0$, we are calculating what has been called the “relative hazard” (see Golub and Steunenberg 2007; Licht 2011). We can also express these quantities as a percentage change in the hazard rate between two values of x at a time (t):

$$\% \Delta h_i(t) = (e^{\beta(x_j - x_l)} - 1) * 100. \quad (3)$$

This is referred to as the first difference.

So far we have only considered linear additive relationships. We can easily extend these quantities of interest to express interactive and nonlinear effects.

Polynomial Nonlinear The n th degree polynomial nonlinear effect is given by $\beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_n x_i^n$. So the hazard rate for units x_j and x_l would simply be:

$$\frac{h_j(t)}{h_l(t)} = e^{(\beta_1 x_{j-l} + \beta_2 x_{j-l}^2 + \dots + \beta_n x_{j-l}^n)}. \quad (4)$$

where $x_{j-l} = x_j - x_l$. There are analogous relative hazards and first differences.

Penalized Splines Another useful and more flexible way of modeling nonlinearity in Cox PH models are penalized splines (see [Gray 1992](#); [Keele 2008](#)). Penalized splines (P-splines) are essentially “linear combinations of B-spline basis functions” ([Strasak, Lang, Kneib, Brant, Klenk, Hilbe, Oberaigner, Ruttman, Kaltenbach, Concini, Diem, Pfeiffer, and Ulmer 2009](#), 5) with joins at observed values of x known as “knots” (k) ([Keele 2008](#), 50). The knots are equally spaced over the range of observed x . If $g(x)$ is the P-spline function then a Cox PH model with P-splines is given by:

$$h(t|\mathbf{X}_i) = h_0(t)e^{g(x)}. \quad (5)$$

For the purposes of post-estimation simulations $g(x)$ is a series of linear combined coefficients:

$$g(x) = \beta_{k_1}(x)_{1+} + \beta_{k_2}(x)_{2+} + \beta_{k_3}(x)_{3+} + \dots + \beta_{k_n}(x)_{n+}, \quad (6)$$

where n is the number of knots. x_{c+} for a given β_{k_c} is:

$$(x)_{c+} = \begin{cases} x & \text{if } k_{c-1} < x \leq k_c \\ x & \text{if } x \leq k_1 \text{ and } k_c = k_1 \\ x & \text{if } x \geq k_n \text{ and } k_c = k_n \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where x is within the observed range data. So, the hazard ratio between x_j and x_l :²

$$\frac{h_j(t)}{h_l(t)} = e^{g(x_j) - g(x_l)}. \quad (8)$$

We can similarly find first differences and relative hazards.

Linear Multiplicative Interactions A linear multiplicative interaction between two variables x and z for unit i is the combined effect of $\beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$. We can easily express this as a hazard ratio:

$$\frac{h_j(t)}{h_l(t)} = e^{(\beta_1 x_{j-l} + \beta_2 z_{j-l} + \beta_3 x_{j-l} z_{j-l})}, \quad (9)$$

Again relative hazards and first differences can be easily found by extension.

[Brambor, Clark, and Golder \(2006\)](#) argue that multiplicative interactions are more easily communicated as marginal effects. For Cox PH models the marginal effect of x (ME_x) on the hazard rate for different values of z_i is given by:

$$ME_x = e^{(\beta_1 + \beta_3 z_i)}, \quad (10)$$

using the same notation as the previous equation.

²For an extension with time-varying effects see [Strasak et al. \(2009\)](#).

Time Interactions In cases where the effect of a covariate x on the hazard rate changes over time, it can be useful to explicitly model this as an interaction between x and some function of time [Licht \(2011\)](#); [Box-Steffensmeier, Reiter, and Zorn \(2003\)](#); [Box-Steffensmeier and Jones \(2004\)](#). If $f(t)$ is a function of time, such as log-time, then a Cox PH model with one time-interaction is represented by:

$$h_i(t|\mathbf{x}_i) = h_0(t)e^{(\beta_1 x_i + \beta_2 f(t)x_i)}. \quad (11)$$

A hazard ration between units j and l is given by:

$$\frac{h_j(t)}{h_l(t)} = e^{(x_j - x_l)(\beta_1 + \beta_2 f(t))}, \quad (12)$$

with obvious relative hazard and first difference extensions.

1.2. Draws from the Posterior Distribution

To explore and communicate the uncertainty we have about quantities of interest calculated from point estimates $\hat{\beta}$ of the Cox PH model we can simulate values of these quantities of interest. One way to do this is by drawing estimates of the coefficients $\hat{\beta}$ from the multivariate normal distribution with a mean of the estimated parameter and parameter-covariance estimates ([King et al. 2000](#); [Licht 2011](#)).³ This a relatively easy way to find information about the parameters' probability distribution.

Once we have the simulated parameters we can use the appropriate formula discussed above to calculate the quantities of interest for each pair of x_j and x_l and, in the case of time interactions, each time t . We can then graphically communicate the distribution of the simulated quantities of interest by, for example, simply plotting the simulated values as points on a figure. This will give you an your reader an easy to read way of understanding your results.

1.3. Which Uncertainty Interval: The Central Interval or the Shortest Probability Interval

Before illustrating how **simPH** can be used to simulate and graph quantities of interest. Let's look at two issues related to how to communicating the quantities' probability distribution. The first issue is what interval should we focus on; some constricted central interval, e.g. the middle 95 percent of the simulations or the highest density region? In the next subsection we will look at how to use visual-weighting to emphasis parts of the distribution with the highest probability.

Many researchers choose to communicate uncertainty about their estimates and quantities of interest using standard errors or confidence intervals. [Licht \(2011\)](#) chose to graphically communicate the uncertainty about the simulated time interactions from her models with lines delimiting the simulated quantity of interest's 2.5 and 97.5 percentiles: the central 95 percent interval.

Another approach to showing uncertainty from a simulated distribution is to use the highest density regions ([Box and Tiao 1973](#); [Hyndman 1996](#)). These are the parts of the distribution

³[King et al. \(2000\)](#) discuss alternatives approaches to finding similar information such as fully Bayesian Markov-Chain Monte Carlo estimation and bootstrapping. They are all similar and differ largely in the way the parameters are draw.

with the highest concentration of a given percentage of simulations. If the simulations are unimodal, [Liu, Gelman, and Zheng \(2013\)](#) recommend finding the shortest probability interval (SPIn) or the shortest interval with a given probability coverage.

If the simulations are normally distributed, the central interval and the SPIn will be equivalent. However, if the distribution is bounded asymmetric then the SPIn is preferable because the central interval “can be much longer and have the awkward property [of] cutting off a narrow high-posterior slice that happens to be near the the boundary, thus ruling out a part of the distribution that is actually strongly supported bu the inference” ([Liu et al. 2013](#), 2).

This is important for us because the quantities of interest we are estimating are on an exponential scale and bounded. Hazard ratios and relative hazards are bounded at 0 and first differences are bounded at -100. So in many cases the SPIn will be a more appropriate way of showing the likely values we can infer from our analysis.

1.4. Visual Weighting

Whether representing a given central or shortest probability interval, it is common to use lines to represent the edges on a graph. The only other information given to the reader is typically a third line for some measure of central tendency. Some graphs shade the region between the interval’s edges. This approach over emphasizes the edges, the areas of lowest probability. Uniform shading suggest to the reader a uniform distribution between the edges. Both of these characteristics give misleading information about the quantities of interest probability distribution.

Visual weighting presents a solution to these problems. Hsiang calls visual weight “the amount of a viewer’s attention that a graphical object or display region attracts, based on visual properties” ([Hsiang 2012](#), 3). More visual weight can be created with more “graphical ink” ([Tufte 2001](#)). Visual weight is decreased by removing graphical ink. The simplest way to automatically increase or decrease graphical ink with our simulations is to simply plot each point with some transparency. Areas of the distribution with many simulations will be darker. Areas with fewer simulations, often near the edges will be lighter.

2. The simPH Process

The **simPH** package makes it so that in three steps users can simulate quantities of interest from Cox PH models and graph the results in the ways described above.

- Use **survival**’s `coxph` command to estimate a Cox PH model.
- Simulate parameters estimates, calculate quantities of interest, and keep simulations in a specified interval with **simPH**’s simulation commands.
- Plot the results with **simPH**’s plotting method `simGG`.

As we will see you can add further aesthetic attributes to your `simGG` plots using `ggplot2`.

3. Examples

The following examples illustrate **simPH**’s capabilities for effects estimated using linear coef-

ficients, polynomials, P-splines, linear multiplicative interactions, and time interactions using data from epidemiology and political science.

3.1. Linear Effects

Our first example illustrating how to simulate and plot a linear effect is based on data from the University of Massachusetts AIDS Research Unit IMPACT Study (UIS) (see [Hosmer Jr., Lemeshow, and May 2008](#), 10). The data is accessible in CSV format via UCLA's Institute for Digital Research and Education.⁴ The initial Cox PH model is from [Hosmer Jr. et al. \(2008\)](#) and examples compiled by the [Institute for Digital Research and Education \(2013\)](#). The study looked at the effects of randomly assigned drug treatment programs on the time it took for patients to return to drug use.

Let's first set up and run the model using the **survival** package.

```
# Load survival package
library(survival)

# Download data
uis <- read.csv("http://www.ats.ucla.edu/stat/r/examples/asa/uis.csv")

# Clean up variables for analysis
attach(uis)
drug <- (ivhx == 1)
agec <- age - 30
ndrugtxc <- ndrugtx - 3

# Estimate the model
M1 <- coxph(Surv(time, censor) ~ treat +
agec + drug + ndrugtxc,
method="breslow")
detach(uis)
```

For this example we'll focus on the variable `agec`. It is the subjects' age at their time of enrollment in the study. We centered it at age 30. When we summarize the results we can see that age is estimated to have a negative relationship with return to drug use.

```
summary(M1)

## Call:
## coxph(formula = Surv(time, censor) ~ treat + agec + drug + ndrugtxc,
##       method = "breslow")
##
## n= 593, number of events= 481
## (35 observations deleted due to missingness)
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
```

⁴<http://www.ats.ucla.edu/stat/r/examples/asa/uis.csv>

```
## treat      -0.22710    0.79684    0.09158 -2.48    0.01314 *
## agec       -0.03074    0.96972    0.00794 -3.87    0.00011 ***
## drugTRUE   -0.34258    0.70993    0.10426 -3.29    0.00102 **
## ndrughtxc  0.03091    1.03139    0.00799  3.87    0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## treat          0.797      1.25      0.666      0.954
## agec           0.970      1.03      0.955      0.985
## drugTRUE       0.710      1.41      0.579      0.871
## ndrughtxc      1.031      0.97      1.015      1.048
##
## Concordance= 0.589 (se = 0.014 )
## Rsquare= 0.067 (max possible= 1 )
## Likelihood ratio test= 41.1 on 4 df,  p=2.57e-08
## Wald test              = 43.2 on 4 df,  p=9.26e-09
## Score (logrank) test = 43.4 on 4 df,  p=8.38e-09
```

Now let's use **simPH** to help us get a sense of the magnitude of the estimated effect and our uncertainty about the estimates. Zero has a meaningful value for **agec**, i.e. it is age 30, so the relative hazard has a meaningful interpretation. It represents the ratio of the hazards between someone who entered the trial at an age other than 30 and someone who entered at 30. First use **coxsimLinear** to simulated the relative hazard for fitted values of **agec** between -10 and 26 (the range of values):⁵

```
# Load simPH
library(simPH)

# Simulate relative hazards for agec
Sim1 <- coxsimLinear(M1, b = "agec",
                    Xj = seq(-10, 26, by = 0.5))
```

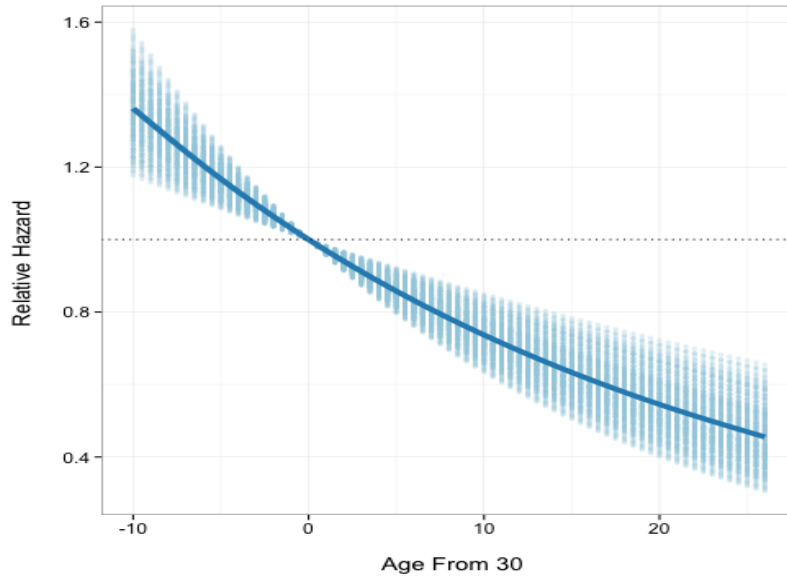
By default, **coxsimPH** finds the relative hazard and the central 95 percent interval of the simulations. Now we plot the results to create Figure 1:

```
simGG(Sim1, xlab = "\n Age From 30")
```

A generalized additive model with integrated smoothness estimation (see [Wood 2013](#)) was automatically chosen to create a line that summarizes the simulations' central tendency across the range of fitted values. This can be changed in **simGG** with the **smoother** argument to any method allowed by **ggplot2**.

If we would instead like to present the results as the percentage change in the hazard between a given age and 30, we could set the **coxsimLinear** **qi** argument to 'First Difference' as in Figure 2.

⁵In R we can use the **seq** command to create a vector with a given sequence of values. In this example we create a sequence with increments of 0.5. The smaller the increments relative to the width of the interval, the smoother the graph will look, though the more increments will require more computing power to simulate.

Figure 1: Simulated relative hazards for Fitted Values of `agec`, Central 95% Interval

```
# Simulate relative hazards for agec
Sim2 <- coxsimLinear(M1, b = "agec",
  qi = "First Difference",
  Xj = seq(-12, 26, by = 0.5))

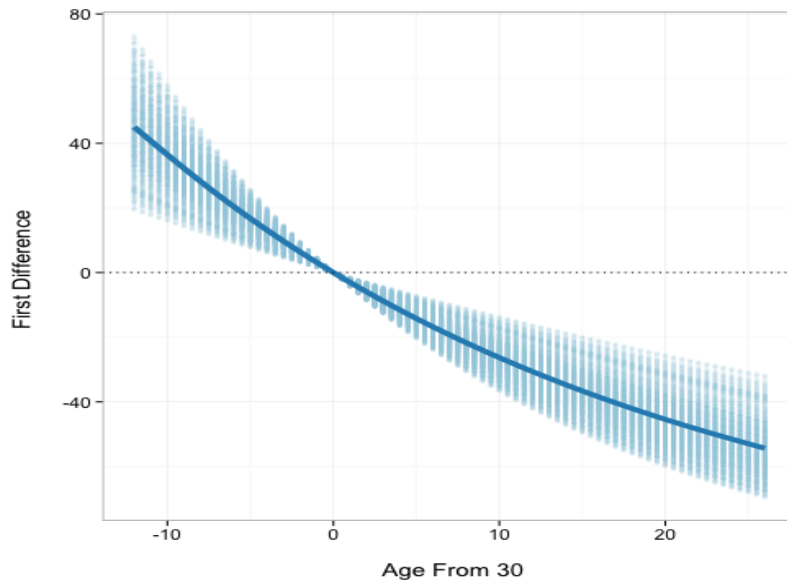
# Plot the results
simGG(Sim2, xlab = "\n Age From 30")
```

4. Conclusion

NOTE: This vignette will be completed in coming versions of **simPH**.

References

- Box GEP, Tiao GC (1973). *Bayesian Inference in Statistical Analysis*. Wiley Classics, New York.
- Box-Steffensmeier JM, Jones BS (2004). *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press, Cambridge.
- Box-Steffensmeier JM, Reiter D, Zorn CJ (2003). “Nonproportional Hazards and Event History Analysis in International Relations.” *Journal of Conflict Resolution*, **47**(1), 33–53.
- Brambor T, Clark WR, Golder M (2006). “Understanding Interaction Models: Improving Empirical Analyses.” *Political Analysis*, **14**(1), 63–82.

Figure 2: Simulated First Differences for Fitted Values of `agec`, Central 95% Interval

- Cox D (1972). “Regression Models and Life Tables.” *Journal of the Royal Statistical Society Series B*, **34**(2), 187–220.
- Gandrud C (2013). *simPH: Tools for simulating and plotting quantities of interest estimated from Cox Proportional Hazards models*. R package version 0.7, URL <http://christophergandrud.github.com/simPH/>.
- Golub J, Steunenberg B (2007). “How Time Affects EU Decision-Making.” *European Union Politics*, **8**(4), 555–566.
- Gray RJ (1992). “Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis.” *Journal of the American Statistical Association*, **87**(420), 942–951.
- Hosmer Jr DW, Lemeshow S, May S (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 2nd edition. Wiley-Interscience, Hoboken, New Jersey.
- Hsiang SM (2012). “Visually-Weighted Regression.” Available at: https://dl.dropboxusercontent.com/u/3011470/WorkingPapers/HSIANG_VISUALLY_WEIGHTED_REGRESSION_v1.pdf.
- Hyndman RJ (1996). “Computing and Graphing Highest Density Regions.” *The American Statistician*, **50**(2), 120–126.
- Institute for Digital Research and Education (2013). “Applied Survival Analysis, Ch. 4: Interpretation of Fitted Proportional Hazards Regression Models.” Accessed May 2013, URL http://www.ats.ucla.edu/stat/r/examples/asa/asa_ch4_r.htm.
- Keele L (2008). *Semiparametric Regression for the Social Sciences*. John Wiley & Sons, Chichester.

- King G, Tomz M, Wittenberg J (2000). “Making the Most of Statistical Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science*, **44**(2), 347–361.
- Licht AA (2011). “Change Comes with Time: Substantive Interpretation of Nonproportional Hazards in Event History Analysis.” *Political Analysis*, **19**, 227–243.
- Liu Y, Gelman A, Zheng T (2013). “Simulation-efficient Shortest Probability Intervals.” *Arxiv*, pp. 1–22. <http://arxiv.org/pdf/1302.2142v1.pdf>.
- Owen M, Imai K, King G, Lau O (2013). *Zelig: Everyone’s Statistical Software*. R package version 4.1-3, URL <http://CRAN.R-project.org/package=Zelig>.
- Strasak AM, Lang S, Kneib T, Brant LJ, Klenk J, Hilbe W, Oberaigner W, Ruttmann E, Kaltenbach L, Concin H, Diem G, Pfeiffer KP, Ulmer H (2009). “Use of Penalized Splines in Extended Cox-Type Additive Hazard Regression to Flexibly Estimate the Effect of Time-varying Serum Uric Acid on Risk of Cancer Incidence: A Prospective, Population-Based Study in 78,850 Men.” *Annals of Epidemiology*, **19**(1), 15–24.
- Therneau T (2013). *survival: Survival Analysis*. R package version 2.37-4, URL <http://CRAN.R-project.org/package=survival>.
- Tufte ER (2001). *The Visual Display of Quantitative Information*. 2nd edition. Graphics Press, Cheshire, CT.
- Wickham H, Chang W (2013). *ggplot2: An implementation of the Grammar of Graphics*. R package version 0.9.3.1, URL <http://CRAN.R-project.org/package=ggplot2>.
- Wood S (2013). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.7-22.

Affiliation:

Christopher Gandrud
 Department of International Relations
 Yonsei University
 208 Jeongui Hall
 1 Yonseidae-gil
 Wonju, Gangwon-do
 220-710, Republic of Korea
 E-mail: gandrud@yonsei.ac.kr
 URL: <http://christophergandrud.blogspot.com>