Indian Journal of Animal Sciences **82** (11): 1423–1428, November 2012

Confidence intervals obtained from different methods using simulated data and their evaluation through artificial neural network

KARIM NOBARI*[1], ALI ASGHAR ASLAMINEJAD[1], MOHAMMAD REZA NASSIRY [1], MOJTABA TAHMOORESPUR[1] and ALI K ESMAILIZADEH[2]

Ferdowsi University of Mashhad, Mashhad, and Shahid Bahonar University of Kerman, Kerman, Iran.

Received: 19 July 2011 ; Accepted: 21 April 2012

ABSTRACT

Determination of confidence intervals (CI) using different methods at different levels of population size (Ps), marker space (Ms), standard deviation of QTL effect (SDQ), ratio of additive to dominance SD (Rad) and QTL position relative to flanking markers (rpQ) were investigated by simulation. The simulation conducted by F_2 design and analyzed with Haley and Knott (HK) method. Moreover an ANN model trained by backprobagation algorithm obtained to predict CIs of different methods at combinations of simulated parameters. After obtain of best ANN model with optimal adequacy parameters we used the artifitial neural network (ANN) model to prediction of CIs at very large-scale combination of simulated parameters comparing actual simulation study. Bootstrap method had more per cent of acurate intervals but average size of the intervals was very high in more scenarios. 1 LOD support interval and bayesian credible interval resulted to be preferable with high per cent of acurate and small confidence intervals, moreover they weekly affected by parameters such as population size and SD of QTL. This study investigated that we can predict CIs for more combination of simulated parameters using best trained ANN. By this study it is sugestive to consideration of more combinations of simulated parameters using the model obtained by best structured ANN to expanding of original study.

Key words: Artificial Neural Network, F_2 Design, QTL mapping, Regression method

The development of genetic maps of markers based upon DNA polymorphisms is beginning to provide the experimental geneticist and the plant and animal breeder with powerful tools for the study of quantitative genetic variation. The use of markers to detect individual loci responsible for quantitative genetic variation (quantitative trait loci or QTL) provides greater power than segregation analysis without marker information. The use of flanking marker methods has proved to be a powerful tool for the mapping of quantitative trait loci (QTL) in the segregating generations derived from crosses between inbred lines (Haley and Knott 1992).

The Haley-Knott regression method is based on multiple regressions which can be applied using any general statistical package, developed by Haley and Knott (1992). They used the example of mapping in an F_2 population and showed that these regression methods produce very similar results to those obtained using maximum likelihood (Haley and Knott 1992). The Haley–Knott (HK) regression method continues to be a popular approximation to standard interval mapping (IM) of quantitative trait loci (QTL) in experimental crosses (Feenstra *et al.* 2006). Currently, the HK regression method is preferred as a fast approximation to the IM method for estimating model parameters (Feenstra *et al.* 2006).

In addition to the correctly estimated QTL position, the size of the CI (95%) serves as an important criterion in QTL mapping. This study carried out to consider effect of different simulated parameters on confidence intervals (CI) obtained from different methods.

Neural networks have been successfully applied in many cases but there has been relatively little research into application of ANNs in the field of animal breeding (Kominakis *et al.* 2002). The ANN model is used to solve a wide variety of problems in science and engineering, particularly for some areas where the mathematical modeling methods fail (Khazaei *et al.* 2008). An ANN model can predict multiple dependent variables based on multiple independent variables, where as a mathematical model is only able to predict one dependent variable at a time (Zhang *et al.* 2002). The most powerful ability of ANN to solve large-scale complex problems is training or education. The best-known and most commonly used training algorithm is back-propagation (Zhang *et al.* 2002 and Drummond *et al.* 2004).

Here, a comprehensive simulation study was carried out

Present address: [1,2]Department of Animal Science, Faculty of Agriculture (k_nobari_ir@yahoo.com).

to determine the effect of marker spacing, population size, standard deviation of QTL effect, ratio of additive to dominance effect of QTL and QTL location relative to flanking markers on the percent of acurate confidence intervals (PACI) and Mean of ACI obtained using HK regression method. Then an optimal ANN was designed to predict PACI and Mean of ACI. Finally the designed adequate ANN model were used for prediction of PACI and mean of ACI for more combination of simulated parameters (scenarioes).

MATERIAL AND METHODS

Haley and Knott regression method

We assume that $y_i|g_i \sim N(\mu_{gi},\sigma^2)$,where y_i is the phenotype of individual i and g_i is its (unobserved) QTL genotype. Therefore we assumed that the phenotypic data (y_i) in every group of animals (animals with same QTL genotype) has a mean depending on QTL genotype and a residual variance. Then we calculate conditional QTL genotype given marker genotype, $p_{ij}=pr(g_i|M_i)$ where M_i is marker genotype data for individual i. Phenotype of individual i for a given marker data follows a mixture of normal distribution. $E(y_i|M_i)=\sigma_j p_{ij}\mu_j$ where μ_j is the mean of individual's phenotype with jth QTL genotype so the conditional phenotype average given marker data is linear in the μj and might be estimated by linear regression of y_i on p_{ij}, thus here at each position across genome we calculate the p_{ij} and then regress the phenotype on this matrix (Broman and Sen 2009).

Data simulation

An F_2 population derived from crossing between two inbred line each with alternate homozygote genotype in marker loci and QTL, with different population size was simulated. 11 markers on one chromosome with different equal spaces of 5 and 10 CM were simulated. Chromosome length was different corresponding to marker spaces from 50 and 100 respectively with presenting one QTL (between 6^{th} and 7^{th} marker). An F_2 populations with different combination of population sizes (Ps) of 300, 600 and 900, Standard deviation of QTL effect (SDQ) of 0.2, 0.5, 0.8 and with different portion of additive to dominance effects (Rad) of 0.25, 0.5 and 0.75 were simulated. In each combination of simulated parameters (scenario) QTL located between sixth and seventh markers relatively with 0, 0.25, 0.5 of the interval separated from sixth marker. Therefor 162 scenarioes with different combinations of 2 levels of marker spacing (Ms), 3 level of Population size (Ps), 3 level of QTL effect (SDQ) each with 3 different levels of proportion of additive dominance effect (Rad) and 3 level of QTL location relative to adjacent flanking markers (rpQ) were considered. Each scenario was replicated 100 times. In each simulated population the trait value had a normal distribution.

For each individual one chromosome with corresponding length was simulated. The genotypes of markers and QTL was sampled from binomial distribution using haldane mapping function, thus Crossing over between markers and between markers and QTL was simulated using haldane mapping function. Trait value was sampled from normal distribution with corresponding mean according to genotype of QTL and with unexplained standard deviation (unexplained by QTL) equal 1 SD(σ=1).

QTL mapping analysis

Analysis was caried out using Haley-Knott regression method using R/qtl (Broman *et al.* 2003) package. At first carried out a genome scan with a single QTL model for estimating probable QTL position with higher LOD score on the chromosome. The LOD scores calculated as $LOD=(n/2)\log_{10}(RSS_0/RSS_1)$ where n is sample size, RSS_0 is the null residual sum of squares and RSS_1 is model residual sum of squres (the model defined as regression of phenotypes on the conditional QTL genotypes depending on markers genotypes). For estimating intercept (mean), additive effect and dominance effectand corresponding standard error of them we fitted a single QTL model using HK regression method. To deterime the significant treshold of LOD score (α= 0.05), a permutation with 1000 replicate using defined model were done for each replication of a scenario. 1 and 1.5 LOD support intervals calculated as interval in wich LOD score is within 1 and 1.5 units of its maximum on the chromosome. With a priori of being QTL existing any where of chromosome is equal, posterior distribution obtained by rescaled LOD to 10^{LOD} to be a distribution, $f(è|data)=10^{LOD(è)}/\sigma_è 10^{LOD(è)}$. The 95% bayesian credible interval is in which $\sigma_{è\circ I} f(è|data)$e"0.95. 95% and 99% confidence interval using non-parametric bootstrap method calculated. In the non-parametric bootstrap, one sample with replacement created from original data with size of equal to original data, in the new data set some individuals omited and some repeated then Interval mapping performed for the sampled data to estimate QTL location. To create a set of locations for QTL, the process of resampling and estimation of QTL location were repeated for 1000 times. 99% and 95% confidence interval of bootstrap method defined as regions covered by 99% and 95% of the 1000 estimated positions, respectively.

Comparision statistics

Here Accurate Confidence Intervals (ACI) is the estimated confidence interval in the replication with significant (α=0.05) QTL that covers real QTL position on the chromosome. Per cent of Acurate Confidence Intervals (PACI) for each scenario was calculated as,

$$PACI = \frac{\text{Number of ACIs}}{\text{Number of significant replicates in the scenario}}$$

We calculated mean of ACIs size as,

$$\text{Mean of ACI} = \frac{\Sigma\ (\text{Size of ACIs in the scenario})}{\text{Number of ACIs in each scenario}}$$

Artificial neural network modeling and evaluation

For fit an ANN model to predict PACI and mean of ACI for differrent parameters using Ps, Ms, SDQ, Rad and rQp as inputs, we developed an ANN model by backpropagation algorithms. The data set obtained from simulation study were divided into learning and testing data set, then learning data set used to train the ANN and testing data set were used for validation of the trained ANN model. The best of number of input, output, learning rate, momentum coefficient, number of hidden layers, number of hidden neurons, and number of training cycles or epochs were chosen to obtain the optimal ANN. To determine adequacy of the ANN model we used three statistics containing R^2, T and root MSE (RMSE). The T statistics measures the scattering around fitted line using the ANN. It's better when close to 1. The formula of calculating T is as below (Khazaei *et al.* 2008),

$$T=1-\frac{\Sigma_i^n=1\ (X_{m,i}-X_{p,i})^2}{\Sigma_i^n=1\ (X_{m,i}-\bar{X})^2}$$

where n is the number of data set, $\bar{X}$ is the average of X over the n samples, and X_m and X_p are the actual and by ANN model predicted HK efficiency parameters, respectively.

Expanding of original simulation study using the ANN models

For appropraite methods of calculating CI accourding to results from the simulation study, we considered more scenarios. Mean of ACI and PACI for these scenarios were predicted using the trained ANN models. Thus More different levels of simulated parameters were used to create large-scale scenarios. Then the adequate ANN models were used to predict outputs. In the original study the space from different levels of simulated parameters was equal thus it can support to proper the use of the adequate ANN model to expanding the original study.

RESULTS

Confidence intervals

PACI using all methods positively affected by increasing Ps and decreasing rpQ. Table 1 shows average effect of other parameters than Ps on PACIs obtained from different methods. Per cent of accurate 1 LOD support interval affected slightly by SDQ and affected by Ms specially in low level of SDQ. SDQ increased the PACI obtained from 1.5 LOD interval the increase was higher when SDQ increased from 0.2 to 0.5. Ms in high and medium SDQ did not affect PACI of 1.5 LOD interval. Increase of SDQ decresed PACIs obtained from bayesian credible interval, other 2 parameters. PACI in the bootstrap 95% and 99% was not affected by Ms, SDQ and Rad.

As presented in Table 2 mean of ACIs in all methods affected by rpQ and sharply by Ps. CIs was small in larg Ps and when the QTL located on the marker. CIs in 1 LOD interval became high with increase of Rad and Ms, size of the CI was narrowed by increas in SDQ, the trend was same in other methods.

In total CIs of bootstrap method was very large than others, size of CIs in different methods. In the SDQ of 0.8 all methids had small CI, and in the SDQ of 0.2 efficiency of bootstraping method decreased. 1 LOD interval was small in SDQ of 0.2, but BCI was better on the context of 0.5 and 0.8.

ANN models and evaluations

To fit an ANN model for prediction of mean of ACIs we used ANN structure of 4–10–5–5, containing number of

Table 1. Average effect of SDQ, Rad and Ms on PACI in different methods

			SDQ=0.2			SDQ =0.5			SDQ =0.8		
			Rad=0.25	Rad=0.5	Rad=0.75	Rad=0.25	Rad=0.5	Rad=0.75	Rad=0.25	Rad=0.5	Rad=0.75
†	Ms	5	91.15	91.05	91.49	97.89	97.33	98.00	99.00	98.44	98.11
		10	90.87	91.93	92.18	97.67	97.55	97.55	98.11	97.78	97.33
‡	Ms	5	97.30	97.00	97.54	99.56	99.00	99.00	100.00	99.44	99.56
		10	95.63	96.90	95.84	98.89	99.00	99.56	99.22	99.11	99.17
•	Ms	5	93.47	94.03	94.13	91.33	92.00	92.22	82.00	87.22	87.89
		10	92.36	94.18	93.42	92.56	92.55	93.77	86.33	89.22	89.67
*	Ms	5	98.04	99.17	99.33	96.22	96.22	97.56	95.56	95.78	94.44
		10	99.60	98.46	98.23	96.89	96.77	97.44	94.78	95.78	95.00
§	Ms	5	99.89	99.76	100.00	99.11	99.33	99.44	99.11	99.00	98.78
		10	100.00	100.00	99.47	98.78	99.56	99.44	98.67	99.00	99.33

Ms, Marker spce, Rad, ratio of additive to dominance Standard Deviation, SDQ, Standard Deviation of QTL effect, PACI, per cent of acurate confidence interval (per cent of confidence intervals that containing QTL), † PACIs that obtained using 1 LOD support interval, ‡ PACIs that obtained using 1.5 LOD support interval, • PACIs that obtained using baysian credible interval with 95%, *PACIs that obtained using bootstrap method with 95%, § PACIs that obtained using bootstrap method with 99%.

Table 2. Average effect of SDQ, Rad and Ms on mean of ACI in different methods in CM

			SDQ=0.2			SDQ =0.5			SDQ =0.8		
			Rad=0.25	Rad=0.5	Rad=0.75	Rad=0.25	Rad=0.5	Rad=0.75	Rad=0.25	Rad=0.5	Rad=0.75
†	Ms	5	17.02	20.24	22.97	6.06	7.79	7.77	3.84	4.61	4.80
		10	22.23	27.61	28.85	8.04	9.99	10.27	5.22	6.24	6.98
‡	Ms	5	23.56	27.38	29.69	7.45	9.85	9.69	4.58	5.55	5.73
		10	30.72	38.53	39.88	9.87	12.56	12.81	6.24	7.59	8.45
•	Ms	5	20.14	24.00	25.53	4.60	6.70	6.40	2.16	2.97	3.09
		10	29.43	35.55	35.10	6.45	8.69	8.83	3.45	4.54	5.27
*	Ms	5	30.11	34.15	35.44	7.99	11.69	11.25	3.50	4.77	4.87
		10	58.80	63.34	61.75	10.10	15.60	14.25	4.76	6.60	7.33
§	Ms	5	40.44	43.92	44.01	12.60	18.33	17.64	5.03	7.21	7.53
		10	81.75	84.76	82.66	16.53	25.99	23.07	6.65	9.79	10.50

Ms, Marker spce, Rad, ratio of additive to dominance Standard Deviation, SDQ, Standard Deviation of QTL effect, ACI, acurate confidence interval (confidence intervals that containing QTL), †, mean of ACIs that obtained using 1 LOD support interval, ‡, mean of ACIs that obtained using 1.5 LOD support interval, •, mean of ACIs that obtained using baysian credible interval with 95%, *, mean of ACIs that obtained using bootstrap method with 95%, §, mean of ACIs that obtained using bootstrap method with 99%.

Table 3. Adequacy statistics of predicted mean of ACI for different methods using the ANN model

	Statistics	1 LOD	1.5 LOD	BCI	Boot 95%	Boot 99%
Testing data set	R^2	0.963404	0.961792	0.963279	0.978809	0.981348
	T	0.958711	0.954043	0.958959	0.977438	0.979699
	RMSE	0.037659	0.041251	0.039236	0.032525	0.033052
Training data set	R^2	0.98213	0.988745	0.986337	0.988626	0.990224
	T	0.981962	0.988522	0.986154	0.98831	0.989874
	RMSE	0.031381	0.026281	0.028501	0.028919	0.028608
All data set (train+test)	R^2	0.977946	0.982191	0.98104	0.986461	0.988097
	T	0.977502	0.981645	0.980654	0.986235	0.987913
	RMSE	0.033442	0.031658	0.032193	0.030076	0.030047

ACI, acurate confidence interval (confidence intervals that containing QTL), 1 LOD, confidence interval resulted from 1 LOD support interval, 1.5 LOD, confidence interval resulted from 1.5 LOD support interval, BCI, confidence interval resulted from Bayesian Credible Interval, Boot 95%, confidence interval resulted from Bootstraping (α=0.95), Boot 99%, confidence interval resulted from Bootstraping (α=0.99).

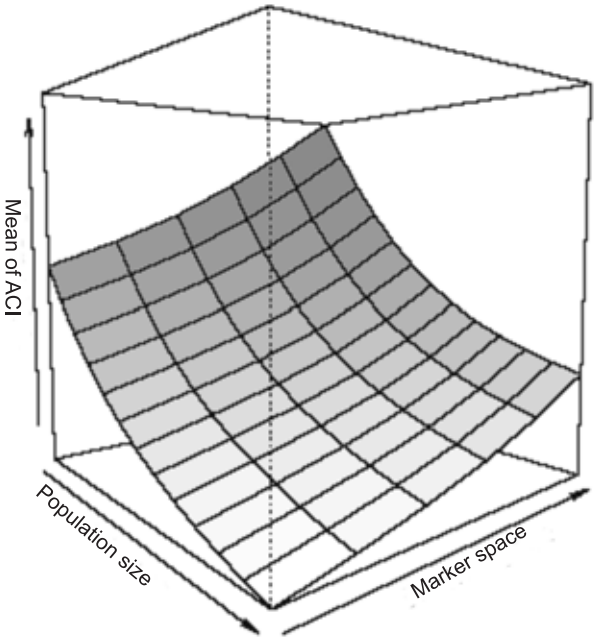
inputs (Ps, Ms, SDQand rpQ), number of neurons in first and second hidden layer and number of outputs, respectively. Results from 162 scenarios partitioned to 110 scenarios for training data set and 52 scenarios for testing data set. Adequacy statistics for every predicted output (mean of ACI resulted from different methods) using the ANN model were presented in Table 3. The ANN trained with Learn Rule of Delta, TanH transfer function, with learning rate and momentum equal to 0.3 and 0.4, respectively. The results show that the ANN model is able to learn the relationship between the inputs and mean of ACIs for different methods.

Adequate ANN structure for prediction of PACI resulted from 1 LOD and BCI methods were 5–10–9–5–2(number of input, numbere of neurons for first, second and third hidden layer and number of output, respectively). Other parameters containing learning rule, transfer function and etc were same with the ANN that fitted for mean of ACI. Adequacy statistic of the ANN model to predict of PACI for 1 LOD and BCI method using the ANN were presented in Table 4. The results show that the ANN can't predict PACI properly, in other word there are other parameter or parameters rather than considered inputs that affects the PACIs.

Expanded study using the ANN models

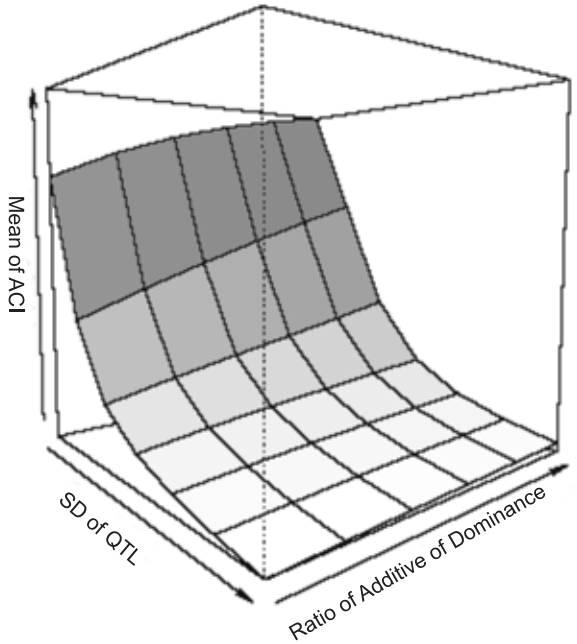
A total of 3,276 different scenarios using different simulated parameters that presented in Table 5 were used to predict mean of ACIs by the trained ANN model. Fig. 1 presents mean of ACIs resulted from 1 LOD support interval in different combinations of Ps and Ms. As presented in the Fig. 1 increase of Ps and decrease of Ms decreases the mean of ACIs. The results show that mean of ACI resulted from BCI affected by Ps and Ms with same trend as 1 LOD support method. Moreover effect of SDQ and Rad on mean of ACIs obtained from 1 LOD support interval were presented in Fig. 2. Effect of the parameters on mean of ACI resulted from BCI followed the same trend as presented in Fig. 2.

Consideration of PACI using the ANN trained for this regard caried out using 36036 different scenarios that



ACI, acurate confidence interval (confidence intervals containing QTL), Ps, Population size, Ms, Marker space

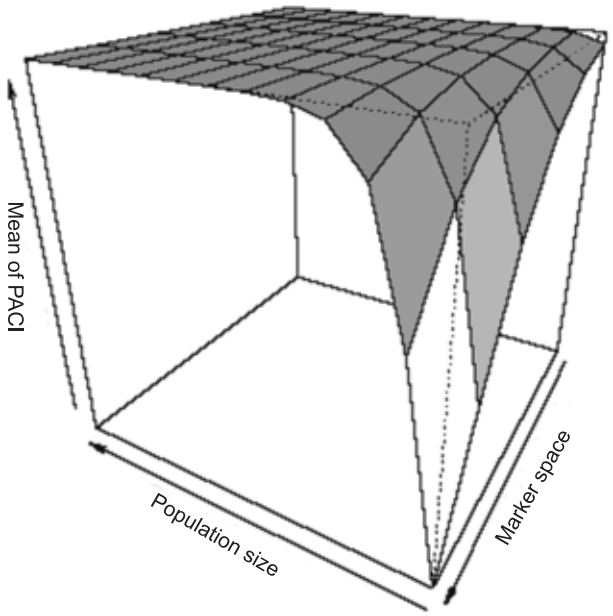
Fig. 1. presents mean of ACIs resulted from 1 LOD support interval in different combinations of Ps and Ms.



SDQ, standard deviation of QTL effect, ACI, acurate confidence interval (confidence intervals that containing QTL)

Fig. 2. Effect of SDQ and Rad on mean of ACIs for 1 LOD support interval.

presented in Table 6. Fig. 3 presents effect of Ps and Ms on different combinations of Ps and Ms on PACI using 1LOD method. Trend of Ps and Ms on PACI of BCI was same as presented in Fig. 3. The result show decreasing of Ms and



Ps, Population size, Ms, Marker space, PACI, per cent of accurate confidence interval (per cent of confidence intervals that containing QTL)

Fig. 3. Effect of Ps and Ms on different combinations of Ps and Ms on PACI using 1LOD method.

Table 4. Adequacy statistics of predicted PACI for 1 LOD support interval and BCI using the ANN model

Data	Statistics	1 LOD	BCI
Testing data set	R^2	0.637163	0.788005
	T	0.313621	0.774682
	RMSE	0.137845	0.087068
Training data set	R^2	0.919026	0.903801
	T	0.918762	0.90253
	RMSE	0.048167	0.054305
All data set(train+test)	R^2	0.777918	0.861341
	T	0.737094	0.860303
	RMSE	0.086376	0.066155

PACI, percent of acurate confidence interval (percent of confidence intervals that containing QTL), 1 LOD, confidence interval resulted from 1 LOD support interval, BCI, confidence interval resulted from Bayesian Credible Interval

increase of Ps increases the per cent of acurate confidence intervals for 1 LOD support and BCI methods. Because of low adequacy of the ANN model trained for PACI no more considerationn caried out.

DISCUSSION

As presented in Table 1 Bootstraping method for confidence interval has a high per cent of acurate confidence interval in all scenarios but mean of confidence interval wide

Table 5. The different combinations of simulated parameters which were analyzed using optimal ANN model for mean of ACI

Population size	Marker space	SD of QTL	Ratio of additive to dominance SD
300	5	0.2	0.25
350	6	0.3	0.35
400	7	0.4	0.45
450	8	0.5	0.55
500	9	0.6	0.65
550	10	0.7	0.75
600		0.8	
650			
700			
750			
800			
850			
900			

ANN, Artificial Neural Network; ACI, acurate confidence interval (confidence intervals that containing QTL); SD, standard deviation.

Table 6. The different combinations of simulated parameters which were analyzed using optimal ANN model for PACI

Population size	Marker space	SD of QTL	Ratio of additive to dominance SD	Relative position of QTL to flanked marker
300	5	0.2	0.25	0
350	6	0.3	0.35	0.05
400	7	0.4	0.45	0.1
450	8	0.5	0.55	0.15
500	9	0.6	0.65	0.2
550	10	0.7	0.75	0.25
600		0.8		0.3
650				0.35
700				0.4
750				0.45
800				0.5
850				
900				

ANN, Artificial Neural Network; PACI, per cent of acurate confidence interval (per cent of confidence intervals that containing QTL) SD, standard deviation.

in all scenarios and only fairly preferable when QTL effect is high. Effect of marker space on size of confidence interval was very small, either Wisscher *et al.* (1996) reported that the Marker spacing had only a small effect on the average empirical confidence interval obtained by Bootstrapping method. Although bayesian credible interval has apropriate per cent of acurate confidence specially in combination of low effect of QTL compairing other methods but this method in the level of effect has higher size of confidence interval comparing 1 lod support interval. Percent of acurate interval of LOD support intervals low affected by parameters, 1 LOD support interval is preferable speciall in high marker space because of pretty per cent of accuracy (Table 1) and size of confidence interval. "Resolving power" defined by Darvasi and Soller (1997) as the 95% of confidence interval fot QTL map location. Their study show that resilving power is inversely proportional to sample size and to square of the QTL gene effect.

The results of this study implies that 1 LOD support interval is better for it's low affection from simulated parameters. This study resulted that ANN can properly learn relationship between simulated parameters and mean of ACI, but it's not usefull for prediction of PACI.

ACKNOWLEDGEMENT

The authors are grateful to Ferdowsi University of Mashhad for supporting the research.

REFERENCES

Broman K W and Sen S. 2009. *A Guide to QTL Mapping with R/qtl*. Springer, New York, USA.

Broman K W, Wu H, Sen S and Churchill G A. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–90.

Darvasi A and Soller M. 1997. A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics* **27**(2): 125–32.

Drummond S T, Sudduth K A, Joshi A, Birrell S J and Kitchen N R. 2004. Statistical and neural methods for site-specific yield prediction. *Trans. ASAE*. **46**: 5–14.

Feenstra B, Skovgaard IM and Broman KW. 2006. Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. *Genetics* **173**: 2269–82.

Haley C and Knott S. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–24.

Khazaei J, Shahbazi F, Massah J, Nikravesh M, and Kianmehr M H. 2008. Evaluation and modeling of physical and physiological damage to wheat seeds under successive impact loadings: mathematical and neural networks modeling. *Crop Science* **48**: 1532–44.

Kominakis A P, Abas Z, Maltaris I and Rogdakis E. 2002. A preliminary study of application of artificial neural network to prediction of milk yield in dairy sheep. Computers and *Electronics in Agriculture* **35**: 35–48.

Zhang Q, Yang S X, Mittal G S and Yi S. 2002. Prediction of performance indices and optimal parameters of rough rice drying using neural networks. *Biosystems Engineering* **83**: 281–90.